Université de Montréal

Recherche d'éléments répétés par analyse des distributions de fréquences d'oligonucléotides

par
Benjamin Provencher

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en informatique

Avril, 2009

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé:

Recherche d'éléments répétés par analyse des distributions de fréquences d'oligonucléotides

présenté par:

Benjamin Provencher

a été évalué par un jury composé des personnes suivantes:

Sylvie Hamel,	président-rapporteur
Miklós Csűrös,	directeur de recherche
Fabian Bastin,	membre du jury

Mémoire accepté le: .....

RÉSUMÉ

Plus de la moitié du génome humain est composé de sous-séquences hautements similaires, répétées à plusieurs endroits. La détection d'éléments répétés est hautement souhaitable puisque d'une part, ils jouent un rôle important au niveaux de l'évolution, et d'autre part, ils faussent et ralentissent les recherches d'homologies.

Ce mémoire présente une nouvelle approche élégante à la recherche d'éléments répétés *de novo*. Cette approche ne nécessite aucune information *à priori* sur la nature des éléments répétés. Nous utilisons plutôt la distribution double Pareto log-normale ajustée à la distribution des fréquences d'oligonucléotides pour repérer les régions répétées, ce qui permet théoriquement d'utiliser l'algorithme sur des séquences non-annotées. Notre approche, qui a été testée sur le chromosome 12 du génome humain ainsi que sur le génome d'autres organismes, atteint une sensibilité moyenne supérieure à 45% tout en conservant une spécificité moyenne supérieure à 80%.

Mots clés: Distribution double Pareto log-normale, éléments répétés, éléments transposables

ABSTRACT

Over half the human genome is made up of highly similar sub-sequences found at many locations. Detecting repeat elements is extremely desirable since they are a driving force of evolution, and furthermore, they skew and slow homology searchs.

This thesis introduces a new refined approach to *de novo* repeat elements search. This approach does not need any *a priori* information on the repeat elements inherent attributes. We rather use double Pareto-lognormal distributions fitted on frequency distributions of genomic words to search for repeated regions. This feature theoretically allows us to use the algorithm on non-annotated sequences. Our approach, who has been tested on human chromosome 12 and on genome of other species, achieves an average sensibility of over 45% while keeping an average specificity of over 80%.

Keywords: Double Pareto-lognormal distribution, repeat elements, transposable elements.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES

ADN	Acide désoxyribonucléique
ADNc	ADN complémentaire
ARN	Acide ribonucléique
DPLN	Double Pareto log-normale
HMM	Hidden Markov model, *Modèle de Markov caché*
ITR	Inverted Terminal Repeat, *Séquence terminale répétée inversée*
LINE	Long Interspersed Nuclear Elements, *Long éléments dispersés*
LTR	Long Terminal Repeats, *Longues répétitions terminales*
ORF	Open Reading Frame, *Phase ouverte de lecture*
RE	Repeated Element, *Éléments Répétés*
SINE	Short Interspersed Nuclear Elements, *Courts éléments dispersés*
TE	Transposable Element, *Éléments Transposables*
UTR	UnTranslated Region, *Région non traduite*

À mon frère et modèle, Mathieu.

REMERCIEMENTS

J'aimerais sincérement remercier mon directeur de recherche Miklós Csűrös pour avoir accepté de me transmettre une partie de son savoir en m'accordant généreusement de nombreuses heures de discussions enrichissantes. Ses cours instructifs ont su développer mon intéret pour cette discipline académique.

J'aimerais remercier ma famille pour leur support et leur confiance, ma copine pour sa patience et sa compréhension exemplaire. J'aimerais remercier mon ami Jean-Sébastien pour avoir réveillé mon intéret pour la biologie. Finalement, je remercie mon ami Jérémie pour ses discussions et réflexions académiques fort enrichissantes.

INTRODUCTION

L'ubiquité des éléments répétés dans le génome des organismes pose de nombreux problèmes pratiques pour la recherche d'homologie, le séquençage et l'assemblage de génomes. Les éléments répétés sont un moteur de l'évolution puisqu'ils influencent la structure et la taille des génomes et créent, modifient et régulent les gènes. Ces propriétés font des éléments répétés un intéressant sujet de recherche d'où la vaste littérature sur leur structure et leur rôle fonctionnel et évolutif.

Le programme le plus connu et utilisé est, sans aucun doute, RepeatMasker [73]. Cette application utilise une librairie d'éléments répétés manuellement compilée pour effectuer une recherche d'homologie sur les séquences cibles. Or, puisque une partie des éléments répétés est souvent spécifique aux espèces, des librairies doivent être compilées à priori avant de pouvoir analyser le génome d'espèces nouvellement séquencées. Puisque de plus en plus de génomes d'espèces distinctes sont séquencés, il est souhaitable de développer des applications de recherche d'éléments répétés *De Novo*, c.-à-d. des applications ne nécessitant pas d'informations à priori. Nous présentons, dans ce présent travail, une application de recherche d'éléments répétés *De Novo* qui utilise la distribution double Pareto log-normale ajustée à la distribution des fréquences d'oligonucléotides.

Le chapitre 1 détaille explicitement les contributions personnelles apportées au présent travail. Dans le chapitre 2, les régions répétées sont abordées d'un point de vue biologique. On y décrit les différentes classes d'éléments répétés ainsi que les motivations derrière leur recherche. Le chapitre 3 traite des séquences biologiques, des mots et des modèles de séquences aléatoires d'un point de vue statistique. L'idée principale et les concepts clés de notre technique de recherche sont énoncés dans le chapitre 4 alors que l'implémentation de l'algorithme est couverte dans le chapitre 5. Finalement, on retrouve dans le chapitre 6 les résultats des tests effectués.

CHAPITRE 1

CONTRIBUTIONS PERSONNELLES

L'idée principale sous-jacente à l'application qui consiste à utiliser la distribution double Pareto log-normale pour modéliser le spectrum des génomes est issue de mon directeur de recherche, M. Miklós Csűrös. J'ai pour ma part approfondis l'idée en effectuant l'analyse des paramètres des distributions DPLN ajustées aux différents spectrums obtenus en faisant varier la taille des mots (Section 3.4.3.1), la longueur des séquences (Section 3.4.3.2) et le nombre de graines espacées utilisées (Section 3.5). Finalement, j'ai implémenté l'algorithme (Chapitre 5) et effectué l'analyse des résultats (Chapitre 6).

CHAPITRE 2

ELÉMENTS RÉPÉTÉS

2.1 Les éléments repétés

Les éléments répétés(RE) sont des séquences d'ADN de grande similarité, ayant des occurrences réparties dans l'ensemble du génome d'un organisme. Les éléments transposables (TE), courtes séquences d'ADN qui possèdent la faculté de se déplacer ou de se dupliquer dans le génome de l'organisme, à l'intérieur d'une même cellule, forment la principale source d'éléments répétés. Les répétitions en tandem, sous-séquences de deux nucléotides et plus répétées en grand nombre, et les duplications segmentales forment l'autre source d'éléments répétés.

2.1.1 Les éléments transposables

On distingue deux classes d'éléments transposables en fonction de leur mécanisme de transposition, la classe I ou rétrotransposon et la classe II ou transposon qui se déplacent par l'intermédiaire d'ARN et d'ADN respectivement.

Les transposon se déplacent généralement d'une région génomique à une autre par un mécanisme conservateur (non réplicatif) similaire au "couper-coller". Il y a donc excision du site d'origine et intégration dans le site cible. Le transposon est composé d'un noyau qui encode un enzyme appelé transposase, borné par deux séquences terminales répétées inversées (en anglais *inverted terminal repeat*, ou ITR) (Figure 2.1). Cet enzyme s'attache aux extrémités inversément répétés du transposon qui peuvent ainsi se lier et créer une structure tige-boucle stable. La transposase clive ensuite le transposon et ligature les extrémités libres de la séquence d'ADN. Le complexe transposon/transposase maintenant libre se lie ensuite à un motif spécifique ailleurs dans le génome où la transposase effectue une coupe décalée dans la séquence cible. Les deux extrémités du transposon sont ensuite fusionnées aux extrémités de la séquence cible par l'ADN polymérase, ce qui a pour effet de créer deux régions directement répétés aux extrémités du tranposon, dans la séquence cible (Figure 2.2). Puisque les sites de restriction de la séquence cible sont peu spécifiques (souvent seulement composés de deux nucléotides), la transposition peut avoir lieu dans plusieurs sites génomiques. Les transposons occupent envi-

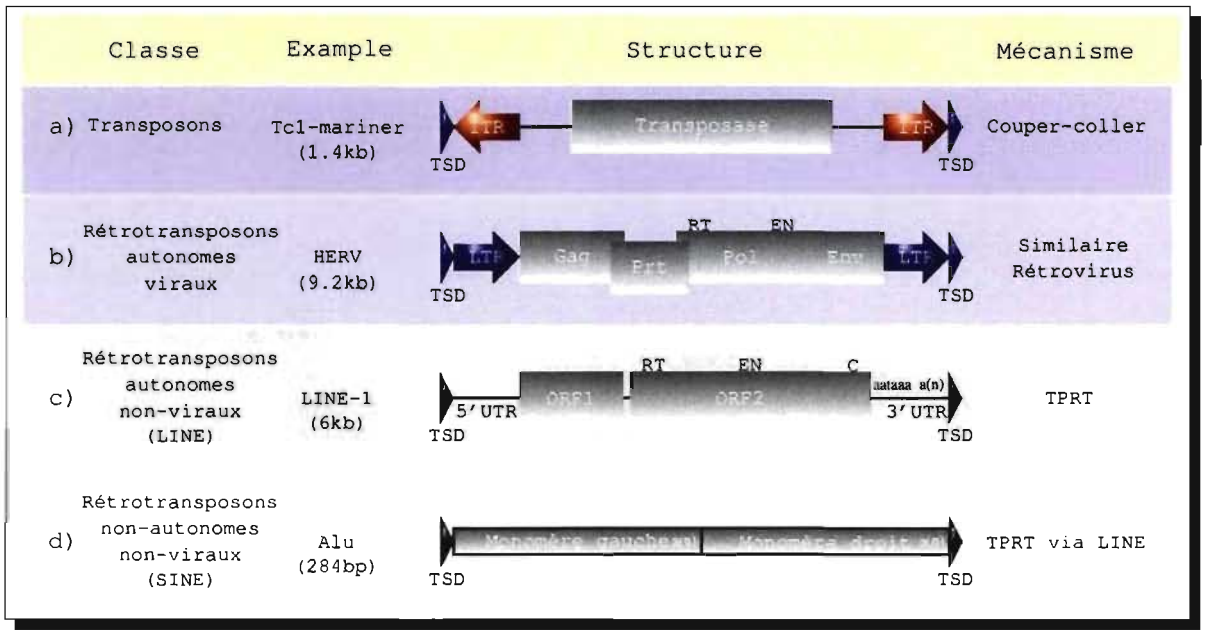


Figure 2.1: Différentes classes d'éléments génétiques transposables chez les mammifères. Chacun des éléments est borné par deux duplications du site cible (TSD) créées pendant le processus d'intégration. (a) Tc1-mariner membre de la famille des transposons d'ADN. Le transposons possède un seul cadre ouvert de lecture, (en anglais *Open Reading Frame*, ou ORF) qui encode la transposase. La transposase est bornée par deux séquences terminales répétées inversées (en anglais *inverted terminal repeat*, ou ITR). (b) Rétrovirus endogène humain (en anglais *human endogenous retroviruses*, ou HERV) membre de la famille des Rétrotransposon LTR. Comme leur nom l'indique, les Rétrotransposons LTR sont bornés par deux longues répétitions terminales (en anglais *long terminal repeats*, ou LTR). Ils sont composés d'ORFs partiellement chevauchants qui encodent les gènes *Gag* (*group specific antigen*), *Prt* (*protéase*), *Pol* (*polymérase*) et *Env* (*envelope*). Sont aussi affichés les domaines de la transcriptase inverse (en anglais *reverse transcriptase*, ou RT) et de l'endonucléase (EN). (c) Élément LINE-1 ou L1 (*long interspersed repeats*) membre de la famille des rétrotransposons non viraux autonomes. Les éléments L1 sont composés de deux ORFs séparés par une courte région intergénique. Le rôle du ORF1 demeure mal compris alors que le ORF2 contient la transcriptase inverse (RT), l'endonucléase (EN) et un motif conservé riche en cystéine (C). Les ORFs sont bornés par une 5'UTR (*UnTranslated Region*) contenant une séquence promotrice pour l'ARN polymérase II et par une 3'UTR contenant un signal de polyadénylation (aataaa) et une queue polyA (a(n)). (d) Élément Alu membre de la famille des rétrotransposons non viraux non autonomes. Les éléments Alus contiennent deux séquences similaires se terminant par une queue polyA (a(n)), les monomères gauche et droit.

ron 2,84% du génome humain [16](Figure 2.6).

Contrairement aux transposons, les rétrotransposons se déplacent d'une région génomique à une autre par un mécanisme réplicatif. L'ARN polymérase (II ou III) transcrit le rétrotransposon en ARN, alors que la séquence originale est conservée. La copie d'ARN est retranscrite en ADN double-brin par la transcriptase inverse, et l'ADN est subséquemment réinséré dans le génome à une nouvelle position. Il y a donc duplication par un modèle analogue au "copier-coller".

Les rétrotransposons peuvent à nouveau être subdivisés en deux catégories ; les rétrotransposons à longues répétitions terminales (en anglais *long terminal repeat*, ou LTR) ou viraux et les rétrotransposons non-LTR ou non viraux.

Les rétrotransposons viraux s'étendent généralement sur 7 à 9 kilobases et possèdent de longues répétitions terminales (formées de 100 paires de bases à plusieurs kilobases) qui bornent la région interne codante contenant tous les éléments régulatoires nécessaires à leur transcription. Plus spécifiquement, on retrouve les gènes *Gag*, *Prt*, *Pol* et *Env* (Figure 2.1). Le gène *Gag* (*group specific antigen*) encode un ensemble de protéines structurales qui forment une VPL (*virus like particles*), structure à l'intérieur de laquelle a lieu la transcription inverse. Le gène *Prt* (protéase) permet de cliver la polyprotéine *Pol*. Le gène *Pol* (polymérase) encode des fonctions enzymatiques telle que la transcriptase inverse qui permet de copier l'ARN du rétrotransposons en ADN complémentaire (voir Figure 2.3 pour une ébauche du mécanisme de trancription inverse), l'endonucléase qui permet de cliver l'ADN ainsi que l'intégrase qui permet de réinsérer l'ADNc dans le génome de la cellule hôte [29]. La différence notable entre les rétrovirus et les rétrotransposons viraux est que ces derniers sont généralement non infectieux puisqu'ils n'ont pas acquis le gène d'enveloppe cellulaire *env* [51], ou alors, en possèdent une version vétuste. De fait, les rétrotransposons à LTR ne peuvent se déplacer de celulle en cellule. On estime que le génome humain est composé à 8.29% de rétrotransposons viraux [16](Figure 2.6).

On subdivise les rétrotransposons non viraux en deux sous-classes ; SINE (*short interspersed nuclear elements* ou courts éléments dispersés) et LINE (*long interspersed nuclear elements*

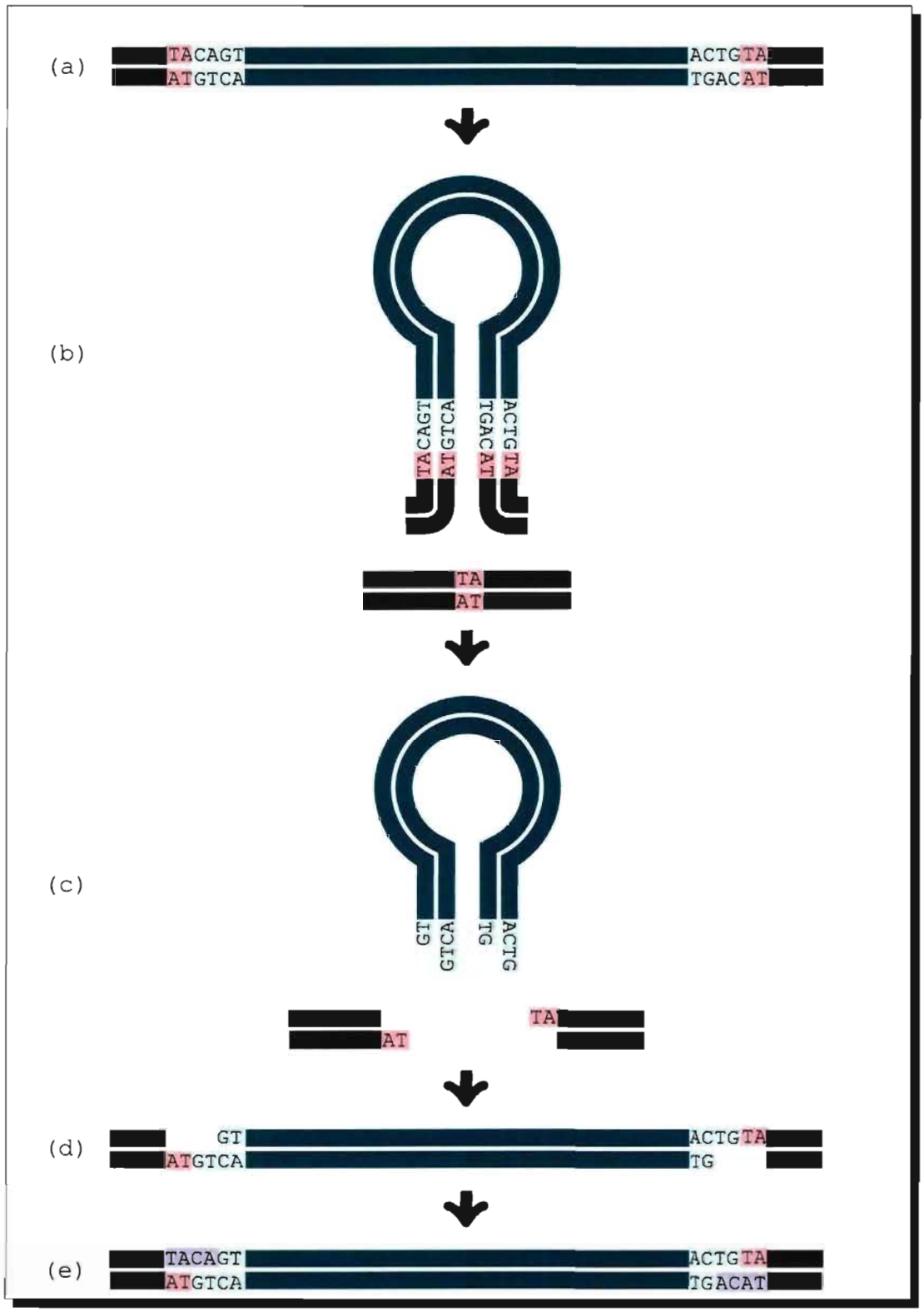


Figure 2.2: Transposition non réplicative. (a) Le transposon borné par les duplications du site cible (TSD, rose) et les séquences terminales répétées inversées (ITR, vert pâle). (b) Sous l'action de la transposase, le transposon adopte une conformation tige-boucle stabilisées par les ITRs. (c) La transposase clive le transposon et ligature les extrémités libres de la séquence hôte d'ADN. Le complexe transposon/transposase maintenant libre se lie ensuite à un motif spécifique (rose) sur la séquence cible où la transposase effectue une coupe décalée. (d) Le transposon est intégré à la séquence cible. (e) Les réparations sont effectuées aux deux extrémités du transposon par l'ADN polymérase (violet). Modifié de [5]

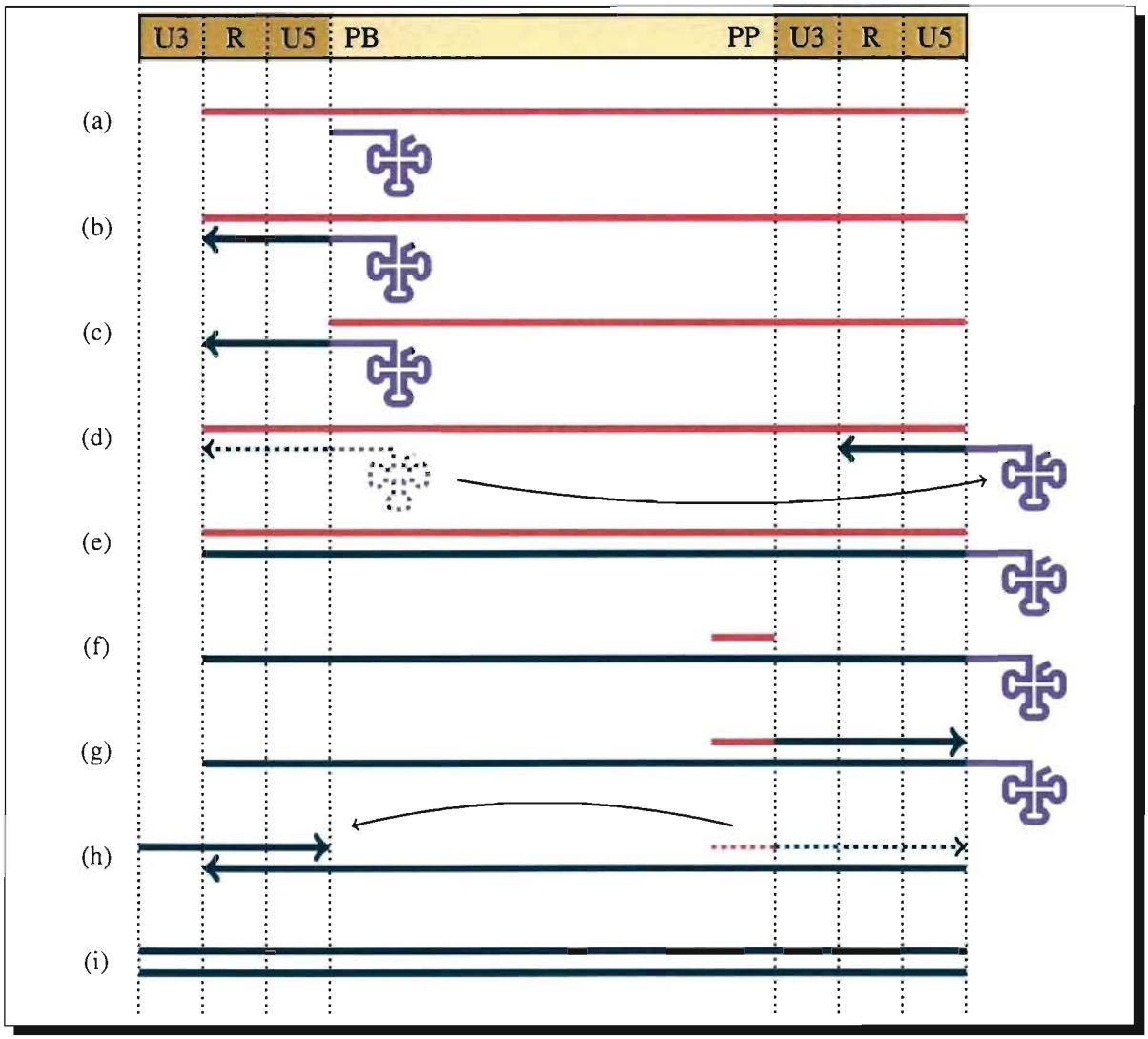


Figure 2.3: Mécanisme de transcription inverse des rétrotransposons viraux. L'ARN messager est dépeint en rose, l'ARN de transfert en mauve et l'ADN en vert. Les lignes flèchées représentent la direction de la synthèse, les lignes pointillées, la position d'un élément avant un saut. (a) La synthèse du brin d'ADN négatif débute à l'extrémité 3'-OH d'un ARN de transfert pairé aux sites de liaison à l'amorce (PB, *Primer Binding Site*) de l'ARN messager. (b) La synthèse du brin d'ADN négatif se poursuit jusquà l'extrémité 5' de l'ARN messager. (c) L'extrémité 5' du brin d'ARNm est dégradée par la ribonucléase H. (d) Premier saut : Puisque Les deux extrémités de l'ARN messager sont identiques, le brin d'ADN négatif peut pairer avec l'extrémité 3'. (e)Après le premier saut, la synthèse de l'ADN se poursuit jusqu'à l'extrémité 5' de l'ARNm. (f) Dégradation de l'ARNm par La ribonucléase H, à l'exception d'une région riche en polypurine (PP, *polypurine tracts*). (g) Cette région est utilisée comme amorce pour la synthèse du brin d'ADN positif, avant d'être à son tour dégradée. (h) Second saut : Le brin d'ADN positif paire avec l'extrémité 5' du brin d'ADN négatif. (i) Lorsque la synthèse des deux brins est terminée, l'ADN peut être réintégré au génome. Modifié de [17]

ou long éléments dispersés).

Les LINEs sont dits "autonomes" puisqu'ils possèdent toute la machinerie nécessaire à leur déplacement. Les 80 à 100 éléments LINEs encore actifs [10] dans le génome humain appartiennent à la grande famille LINE-1 (16.89% du génome humain [16]). Les éléments LINE sont composés de 4 à 6 kilobases et possèdent généralement deux cadres ouverts de lecture (*Open Reading Frame*, ou ORF). Le premier encode une protéine de liaison à l'ARN alors que le second encode les enzymes nécessaires à la rétotransposition autonome, la transcriptase inverse et l'endonucléase (Figure 2.1). Le mécanisme de rétrotransposition des rétrotransposons non viraux demeure mal compris, à l'exception de l'étape de transcription inverse. Cette dernière s'effectue, contrairement aux rétotransposons viraux, grâce à un processus appelé "transcription inverse amorcée sur la cible" (*target primed reverse transposition*) (Figure 2.4). La machinerie de transcription inverse des éléments LINE n'est pas vouée uniquement à la rétrotransposition de leur propre ARN messager (préférence cis) [83] ; elle peut aussi mobiliser des éléments SINE comme les Alus ou d'autres ARN messagers créant ainsi des pseudogènes, et rétrogènes [21, 22, 57]. En conséquence, une grande partie du génome humain a été façonnée, directement ou indirectement, par l'action des éléments LINE ; on estime que 79% des gènes contiennent au moins un fragment d'élément L1 [26].

Les SINEs possèdent typiquement moins de 500 paires de base et n'encodent aucune protéine. De fait, il ne peuvent se retrotransposer par eux-même, ils sont dits "non-autonomes". Les éléments Alus forment la famille de SINE la plus nombreuse dans le génome humain : ils ont émergé il y a 65 millions d'années et se sont amplifiés par rétrotransposition pour atteindre une nombre de plus de un million de copies, comptant pour 13.14% de la masse totale du génome humain [16](Figure 2.6). Différents mécanismes génétiques ont modifié le gène 7SL pour produire l'élément Alu actuel [80](Figure 2.5). Les éléments Alus contiennent deux séquences similaires se terminant par une queue polyA, les monomères gauche et droit (Figure 2.1, 2.5).

2.1.2 Les séquences simples répétées

Le terme "séquence simple répétée" englobe deux types d'éléments répétés, les répétitions en tandem et les duplications segmentales.

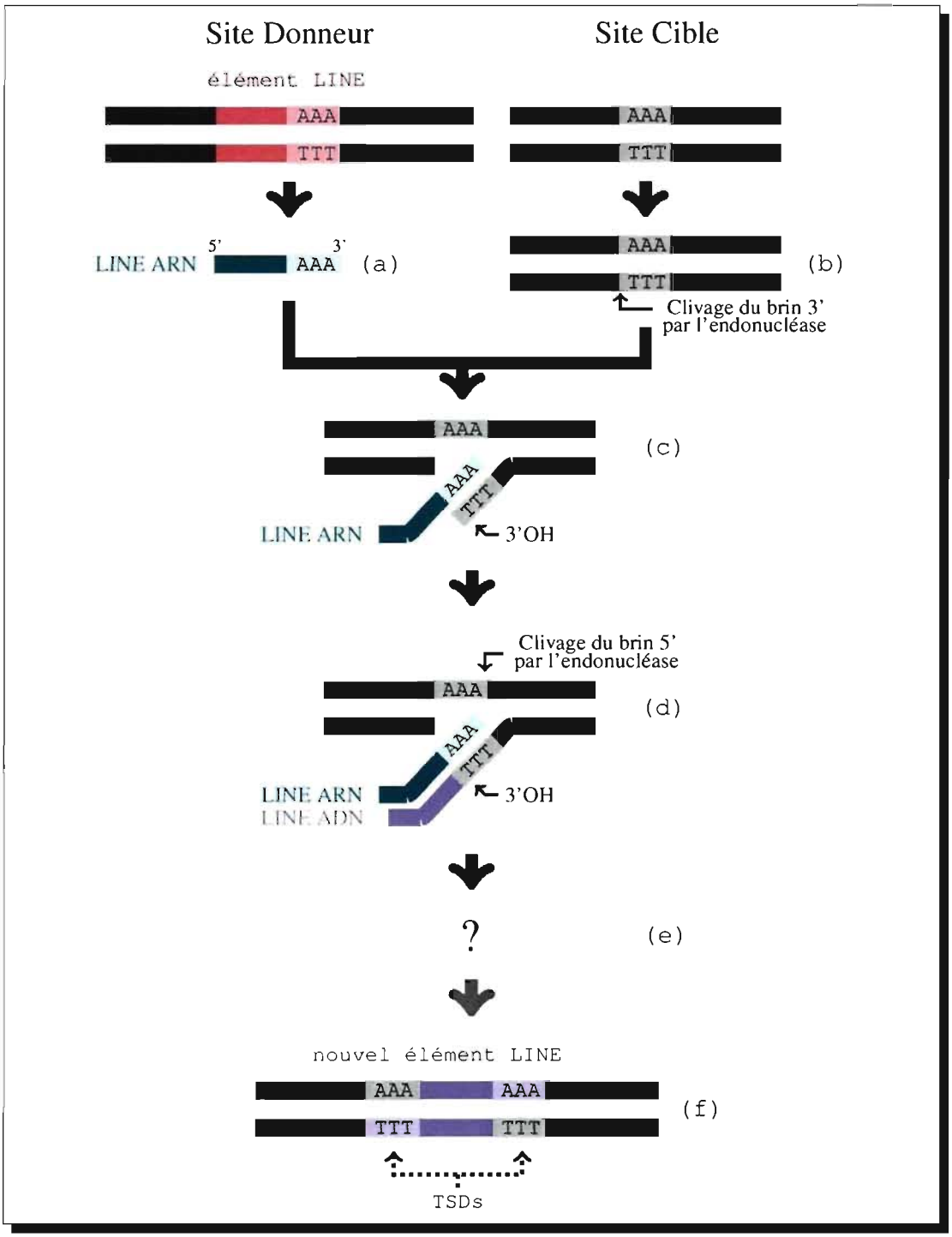


Figure 2.4: Mécanisme de transposition par "transcription inverse amorcée sur la cible". (a) La transposition commence par la transcription de l'élément LINE (rose) en ARN (vert) qui encode l'endonucléase et la transcriptase inverse . (b) L'endonucléase clive ensuite un des brins d'ADN du site cible produisant ainsi un hydroxyle (OH) à l'extrémité 3'OH. (c) La transcription inverse est ensuite effectuée, utilisant l'extrémité 3'OH comme ammorce et le brin d'ARN de l'élément LINE comme modèle. (d) Le clivage du second brin d'ADN cible se produit une fois le brin d'ADN complémentaire (violet) synthétisé. (e) Le mécanisme d'intégration de l'ADNc dans l'ADN cible demeure inconnue. (f) Suppression de l'ARN et synthétisation du second brin d'ADN. L'insertion du nouvel élément LINE produit des duplications du site cible (TSDs) puisqu'il y a duplication de la séquence comprise entre les deux sites de clivage. Modifié de [18]

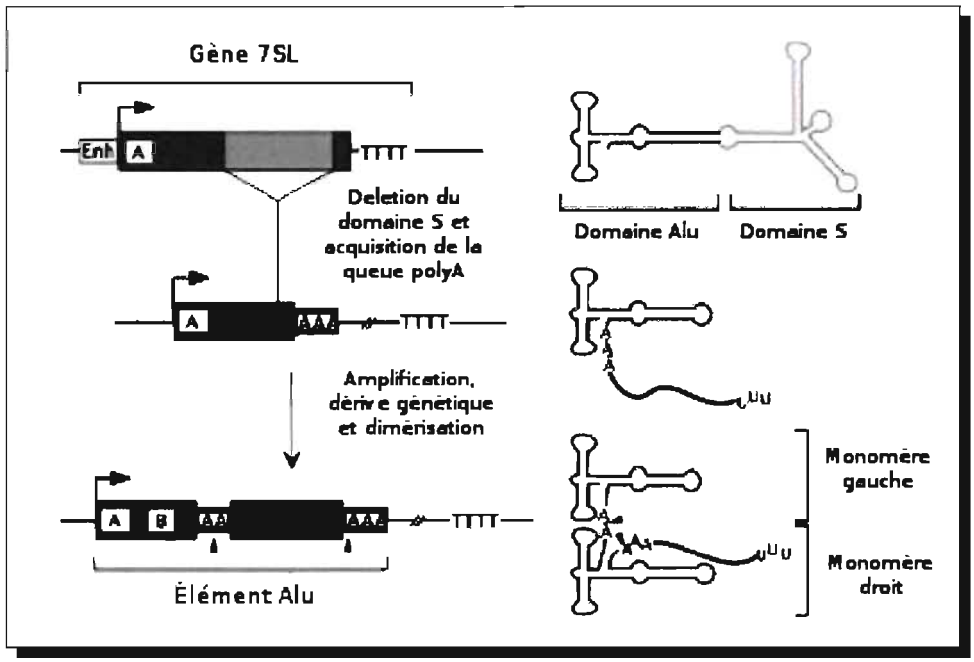


Figure 2.5: Évolution structurale des éléments Alu (gauche) et structure secondaire d'ARN correspondante(droite). L'élément Alu dérive du gène 7SL par un succession d'événements évolutifs : délétions du domaine S, acquisition d'une queue polyA, mutations ponctuelles et duplication tandem. (Tirée de [21])

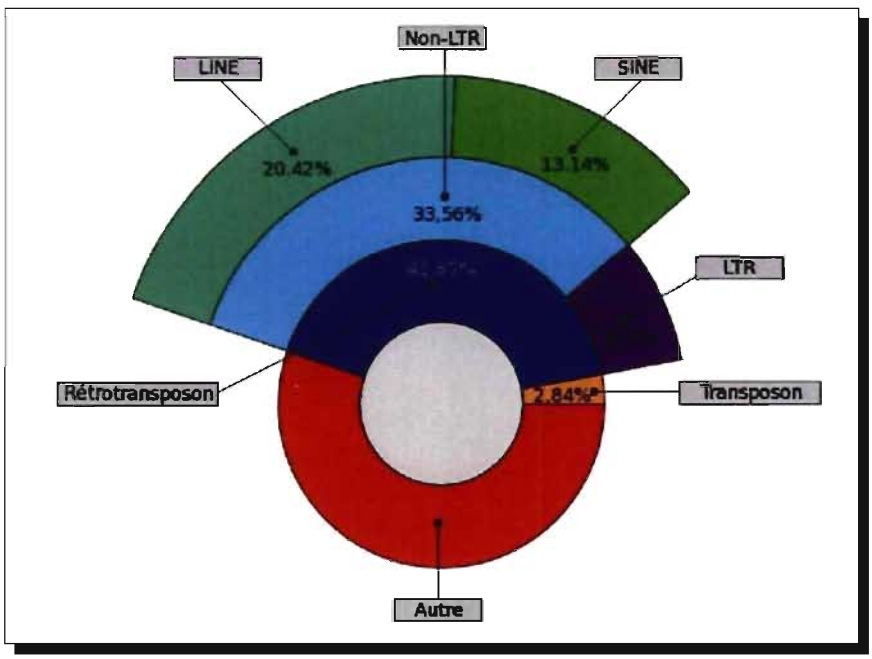


Figure 2.6: Ratio des élément transposables dans le génome humain. (Données provenant de [16])

Les répétitions en tandem sont généralement situées dans l'hétérochromatine constitutive, près des centromères et télomères. Elles sont constituées d'unités primaires reliées les unes à la suite des autres. L'unité primaire s'étend sur une région qui peut faire quelques bases jusqu'à plus d'un kilobase. Les répétitions en tandem consituent les séquences les plus polymorphiques du génome. Contrairement à l'ADN unique, la variabilité se situe au niveau du nombre de répétitions de l'unité primaire plutôt qu'au niveau de la séquence primaire proprement dite. On distingue généralement trois classes de répétitions en tandem en fonction de la taille des unités primaires et de la longueur de la séquence composite : les satellites, les microsatellites et les minisatellites.

Lorsque l'ADN subit une ultracentrifugation dans un gradient de densité, l'ADN se regroupe dans une bande principale visible. Cependant, puisque les répétitions en tandem de courtes sous-séquences modifient la fréquence des nucléotides et, par extension, la densité de l'ADN, on observe souvent une bande satellite distincte de la bande principale, d'où le terme "ADN satellite". L'unité primaire des satellites est composée d'une centaine de paires de bases à plus d'un kilobase. Le satellite entier peut parfois s'étendre sur une région de 100 mb. Plusieurs mécanismes moléculaires seraient à l'origine des satellites ; la recombinaison inégale [74], la réparation de cassure d'ADN double brin [59], la conversion de gènes [24] et la mobilisation par rétrotransposition (par transduction, expliqué plus bas).

Les minisatellites sont composés de sous-séquences répétées de 6 à 100 paires de bases totalisant jusqu'à 30 kilobases, généralement contenues dans les télomères. La génération et la mutation des minisatellites sont généralement expliquées par la recombinaison inégale et la convertion de gènes [4]. On utilise les minisatellites comme marqueurs génétiques dans des applications telles que l'empreinte génétique.

Les microsatellites possèdent une unité primaire s'étendant sur 2 à 6 paires de bases, généralement répétée au maximun une centaine de fois. Puisqu'ils sont intrinsèquement instables, ils sont devenus des marqueurs de choix dans le domaine de la génétique des populations, le taux et la direction des mutations permettant d'estimer la distance génétique entre deux individus. Contrairement aux autres types de satellites, la génération et l'instabilité des microsatellites seraient expliquées par un glissement dans le processus de réplication, c.-à-d. une dissociation

temporaire du brin d'ADN en réplication suivie d'une réassociation désalignée [45]. Les répétitions en tandem couvrent 3% du génome humain [16].

Les duplications segmentales s'étendent sur 1 à 200 kilobases et couvrent 4% du génome humain [86]. Les séquences dupliquées ont généralement plus de 90% d'homologie et peuvent résider sur le même chromosome (duplication intra-chromosomique) ou au contraire, sur un chromosome distinct (duplication inter-chromosomique). Le mécanisme menant aux duplications segmentales demeure nébuleux et fait l'objet d'intenses recherches.

2.2 Les impacts des éléments répétés

On peut diviser les impacts des éléments répétés en trois catégories : les impacts sur le génome (Section 2.2.1), sur les gènes (Section 2.2.2) et finalement, les impacts pratiques (Section 2.2.3).

2.2.1 Les impacts des éléments répétées sur les génomes

Les éléments répétés influencent principalement le génome de trois manières distinctes ; ils modifient la structure(Section 2.2.1.1) et la taille (Section 2.2.1.2) en plus d'être à l'origine de plusieurs réarrangements(Section 2.2.1.3).

2.2.1.1 Structure du génome

Des études ont démontré des similarités entre le mécanisme de rétrotransposition des rétrotransposons non viraux et l'action de la télomerase sur les telomères. Les deux processus peuvent utiliser les extrémités 3'OH des brins d'ADN endommagés à l'intérieur ou a l'extrémité des chromosomes comme ammorce de la transcription inverse [47, 60]. Ainsi, deux rétrotransposons non viraux, *HeT-A* et *TART*, jouent le rôle de la télomérase dans le génome de *Drosophilia melanogaster* [6, 46, 61]. Ces deux rétrotransposons se transposent spécifiquement aux extrémités des chromosomes. La juxtaposition successive des éléments *HeT-A* et *TART* créée des régions répétées qui, quoique plus longues et plus complexes, sont analogues à celles créées par la télomérase dans d'autres organismes. D'autres insectes tel que le ver à soie *Bombyx mori* possèdent aussi des rétrotransposons non viraux, *TRAS1* et *SART1* ,

qui ciblent spécifiquement les télomères [58, 78]. Il a aussi été démontré par Higashiyama *et al.* [31], qu'une famille de rétrotransposons non viraux, les éléments *Zepp*, contenus dans le génome de l'algue *Chlorella vulgaris*, s'intègrent préférablement dans d'autres éléments Zepp lors de leur rétrotransposition. Le groupe suggère que les répétitions en tandem produites par les rétrotranspositions successives pourraient créer et réparer les télomères. Morrish *et al.* [56] propose que les rétrotransposons non viraux sont un mécanisme ancestral de réparation des télomères, avant l'acquisition du domaine de l'endonucléase.

2.2.1.2 Taille du génome

La rétrotransposition d'éléments mobiles engendre évidemment un accroissement de la taille des génomes. Il n'est donc pas surprenant de constater une forte corrélation positive entre la taille des génomes et leur contenu en éléments transposables [37, 49]. À titre d'exemple, la taille du génome du maïs a doublé (passant d'environ 1200 Mb à 2400 Mb) lors des trois derniers millions d'années, sous l'action des éléments répétés [68]. La machinerie de rétrotransposition des éléments LINE permet, outre leur propre mobilité (Figure 2.7 A), de déplacer des éléments transposables non autonomes [21] ainsi que d'autres ARN messagers, engendrant ainsi des pseudogènes [22, 57](Figure 2.7 B). Les pseudogènes engendrés sont généralement non fonctionnels. Cependant, des preuves suggèrent qu'au moins huit pseudogènes du génome humain seraient fonctionnels [9].

Les éléments transposables influent également sur la taille des génomes de manière indirecte par des mécanismes telles la recombinaison homologue, la transduction ainsi que par la duplication de gènes [37].

2.2.1.3 Réarrangement de génomes

En plus de restructurer directement le génome lors de leur insertion, l'homologie des éléments répétés leur permet de servir de substrats aux recombinaisons ectopiques, restructurant ainsi le génome de manière indirecte (Figure 2.7E-G). Les transposons peuvent aussi, lors de leur excision du site cible, créer une cassure éphémère du double brin d'ADN qui peut potentiellement amorcer une recombinaison [77]. On retrouve plusieurs cas de recom-

binaison *in vitro* et *in vivo* dus à différents types d'éléments répétés dans la littérature (e.g. [25, 27, 28, 40, 64, 71, 72, 79, 84]). Dans le seul génome humain, Deininger et al. on trouvé 33 cas de maladies génétiques et 16 cas de cancers causés par une recombinaison entre éléments Alus [42]. Les recombinaisons entre éléments LINE, quoique beaucoup plus rares, peuvent aussi causer diverses pathologies [13, 70].

2.2.2 Les impacts des éléments répétés sur les gènes

Les RE influent sur les gènes en créant de nouveaux gènes ou en modifiant des gènes déja existant (Section 2.2.2.1). Certains REs sont aussi impliqués dans la régulation de gènes (Section 2.2.2.2).

2.2.2.1 Création et modification de gènes

En plus de modifier et de dupliquer les gènes indirectement en servant de substrats aux réarrangements de génomes, les éléments transposables les influencent aussi directement. Puisque la transposition s'effectue généralement sans discernement par rapport au site d'insertion, il est possible que des éléments s'insèrent dans les gènes. Une insertion dans un exon mène généralement, au mieux, à un épissage alternatif, ou au pire, à une perturbation du cadre ouvert de lecture causant ainsi la troncature de la protéine (e.g [62], Figure 2.7 C.1). Cependant, une insertion dans un intron peut aussi avoir des conséquence néfastes ; l'élément inséré peut directement contenir un codon d'arrêt, ou encore déplacer le cadre ouvert de lecture, créant ainsi un codon d'arrêt dans un exon en aval (Figure 2.7 C.2). Il a été démontré qu'une seule mutation ponctuelle peut mener à l'*exonization* d'un élément Alu contenu dans une région intronique, étant donné leur grande similarité avec les sites d'épissages [44]. Or, l'insertion d'un exon peut aussi modifier irrémédiablement le transcrit encodant la protéine originale, créant par le fait une désordre génétique. De fait, des pathologies causées par des éléments Alus *exonizés* ont été démontrées [39, 54, 81, 82].

Puisque les éléments L1 possèdent un faible signal de clivage et de polyadénylation, leur transcrit est parfois clivé après un signal de polyadénylation tierce, situé en amont (Figure 2.7D). Ce processus, appelé 3' transduction, permet d'effectuer un "rétro-brassage d'exon" en insérant de nouveaux exons dans un gène donné [55]. En utilisant une approche bio-informatique, Pickeral *et al.* affirment qu'environ 15% des éléments L1 complets contiennent une séquence

"*transductée*" et qu'en extrapolant sur les 600000 copies contenues dans le génome humain, environ 1% du génome proviendrait de telles séquences [63].

Récemment, un intérêt grandissant a été porté à deux classes d'éléments transposables vu leur capacité particulière à capturer des fragments de gènes et à les déplacer à travers le génome. En effet, plusieurs cas de transduction de séquences géniques par une classe de transposon appelée MULE (Mutator-like transposable elements) ont été documentés dans les génomes du riz [34, 36] et de l'arabette des dames (*Arabidopsis thaliana*) [85]. Plus remarquable encore, il a été démontré qu'un gène complet a été dupliqué par un élément transposable de la famille des Helitrons dans le génome du maïs et que ce gène demeure fonctionnel [32].

2.2.2.2 Régulation de gènes

En plus des introns et des exons, les séquences promotrices, les *enhancers* et les *silencers* peuvent être la cible de rétrotransposition, ce qui peut influer sur la régulation des gènes. Une étude a démontré que 24% des 2004 séquences promotrices 5' analysées contiennent un élément transposable alors que 2% des 846 éléments cis-regulatoires analysés dérivent d'un élément répété [35]. Puisque plusieurs rétrotransposons se transcrivent eux-mêmes, ils possèdent souvent leur propre séquence promotrice. Les éléments L1 possèdent une activité promotrice en amont et en aval dans leur 5'UTR. De fait, les éléments L1 complets peuvent promouvoir la transcription des gènes les entourant, menant potentiellement à l'expression ectopique des dits gènes.

2.2.3 Les impacts pratiques des éléments répétés

D'un point de vue pratique, les éléments répétés ont une influence sur l'assemblage de génome (Section 2.2.3.1), l'annotation de génome (Section 2.2.3.2) et sur la recherche d'homologie (Section 2.2.3.3).

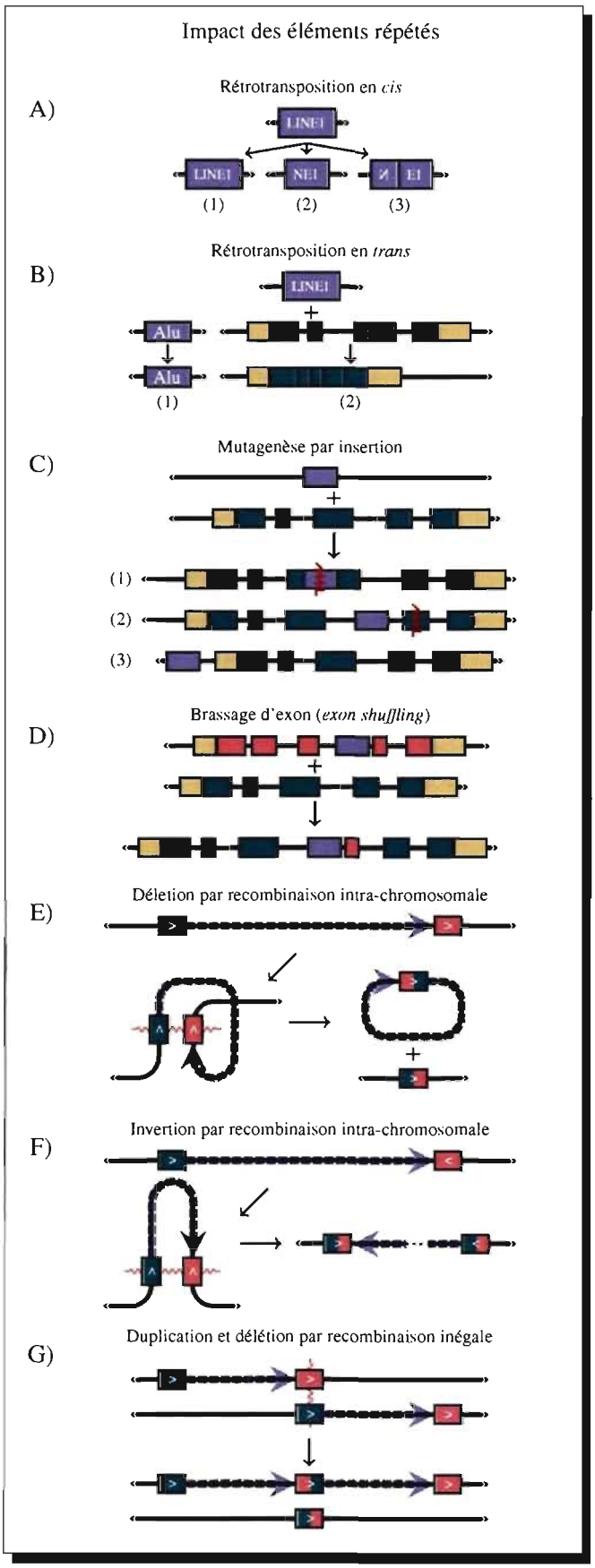


Figure 2.7: Impact des éléments mobiles sur le génome. (a) La rétrotransposition influe sur la taille des génomes. La machinerie de transcription inverse des éléments L1 cible de préférence l'ARN qui l'a traduit (préférence *cis*). Les éléments LINE1 peuvent se retrotransposer entièrement (1), partiellement (2) ou même partiellement inversés (3). b) La machinerie de rétrotransposition des élément LINEs peut aussi être utilisée par les éléments transposables non autonomes tels que les élément alu(1) ou par d'autres ARNs messagers créant ainsi des pseudogènes et rétrogènes (2). c) Mutation génique par insertion d'un élément transposable. L'insertion d'un TE (mauve) dans un exon (vert) mène généralement à la troncature de la protéine suite à l'insertion d'un codon d'arrêt (ligne en rouge en dents de scie) (1). L'insertion dans une région intronique peut aussi être cause de désordre génétique en créant un glissement du cadre ouvert de lecture (2). Les régions promotrices des éléments répétés transposées en amont d'un gène peuvent influer sur la régulation de ce dernier (3). d) Il arrive occasionnellement qu'une sous-séquence située en amont d'un élément L1 (mauve) soit aussi mobilisée lors de la transposition, phénomène appelé "transduction 3'". Le brassage d'exon peut résulter d'une transduction 3' contenant un exon. e) S'ils sont dans la même direction, sur la même séquence, deux éléments répétés peuvent pairer et se recombiner. La recombinaison intra-chromosomale mène à la déletion de la séquence située entre les deux éléments répétés puisque cette dernière ne possède pas de télomère. f) Si, au contraire, les deux éléments répétés sont en direction inverse, la recombinaison intra-chromosomale mène à l'inversion de la séquence située entre les deux éléments répétés. g) Une recombinaison inégale entre deux éléments répétés mène à une délétion et une duplication.

2.2.3.1 Assemblage de génomes

Le séquençage *shotgun* est une méthode utilisée pour séquencer de longs brins d'ADN. Puisque la synthétisation d'ADN par la méthode de Sanger ne peut être utilisée que sur des fragments relativement courts (100 à 1000 paires de bases), les longs brins d'ADN doivent être fractionnés en plusieurs sous-séquences. Ces sous-séquences sont par la suite synthétisées en courtes séquences appelées *reads*. Plusieurs *reads* chevauchant sont obtenus en fragmentant et sequençant le long brin d'ADN à plusieurs reprises. Les *reads* sont assemblés en utilisant l'information fournie par le chevauchement des extrémités pour obtenir la séquence originale. Il est cependant très difficile d'assembler les régions répétées, surtout si celles-ci sont plus longues que les *reads*.

2.2.3.2 Annotation de génomes

Étant donné leur forte dégénérescence et leur structure imbriquée complexe, l'annotation de TE est un processus ardu qui demande beaucoup de ressources humaines et matérielles. De plus, la présence d'éléments transposables nuit à l'annotation de gènes. Les ORFs et séquences promotrices des rétrotransposons peuvent faire en sorte qu'ils soient confondus par les applications de recherches de gènes. Les pseudogènes, produits des rétrotransposons, peuvent aussi être sources de confusion.

2.2.3.3 Recherche d'homologie

Lors de recherche d'homologie, la séquence recherchée peut être alignée à plusieurs occurrences d'un même TE, mais de tels alignements n'apportent généralement pas d'informations biologiques pertinentes. Puisque le temps de recherche augmente généralement linéairement avec le nombre d'alignements trouvés, la présence d'éléments répétés peut considérablement atténuer la performance des algorithmes.

CHAPITRE 3

SÉQUENCES BIOLOGIQUES, MOTS ET MODÈLES

La section 3.1 de ce chapitre définie les séquences d'ADN du point de vue probabiliste. La section 3.2 présente quelques modèles statistiques de séquences aléatoires généralement utilisés, alors que la section 3.3 aborde la distribution des occurrences de mots. La section 3.4 introduit les spectrums et la distribution double Pareto log-normale utilisée dans notre application. Finalement, la section 3.5 définit les graines espacées.

3.1 Séquences biologiques

Le génome d'un organisme vivant est constitué d'une ou de quelques très longues molécules d'ADN, organisées en chromosomes. L'ADN est une structure linéaire, composée de deux brins, formant une double hélice (Figure 3.1). Chaque brin est composé d'une chaîne d'éléments appelés nucléotides. Un nucléotide est composé d'un groupement phosphate, d'un sucre à 5 atomes de carbones appelé désoxyribose et de l'une des quatre bases azotées possibles : l'adénine, la cytosine, la guanine et la thymine. Les nucléotides sont reliés aux niveaux des positions 3' et 5' des carbones du désoxyribose pour former un brin. En conséquence, chaque brin possède une polarité, puisqu'un brin débute avec un groupement phosphate en 5' et se termine par un groupement OH en 3'. L'hélice d'ADN est composée de deux brins antiparallèles (la polarité des deux brins est inversée), rattachés par des liaisons d'hydrogène entre deux nucléotides complémentaires ; l'adénine s'apparie uniquement avec la thymine (Figure 3.2(a)) et la cytosine, avec la guanine (Figure 3.2(b)). Cette complémentarité de l'hélice d'ADN permet la réplication semi-conservatrice du génome ; chacun des deux brins parentaux sert de modèle à la synthèse d'un brin enfant, créant ainsi deux nouvelles hélices identiques.

On considère généralement un brin d'ADN comme une simple séquence de lettres de l'alphabet $\mathcal{A} = \{a, c, g, t\}$ dénotée par $\mathbf{S} = s_1 \dots s_l$, où l correspond à la longueur de la séquence. Cependant, pour des raisons pratiques, nous définissons plutôt une séquence biologique comme une séquence finie de variables aléatoires $\mathbf{X} = (X_i)_{i=1,\dots,l}$ sur l'alphabet $\mathcal{A}$. Nous sommes aussi intéressés par les mots contenus dans une séquence $\mathbf{X}$. Nous dénotons par $\mathbf{w} = w_1 \dots w_k$, un mot de taille k sur l'alphabet $\mathcal{A}$. La fonction indicatrice d'un mot $\mathbf{w}$ situé à la position i dans $\mathbf{X}$

est donné par

$$Y_{\mathbf{X}}(i,\mathbf{w}) = \begin{cases} 1 & \text{si } X_i \dots X_{i+k-1} = w_1 \dots w_k \\ 0 & \text{sinon} \end{cases} \tag{3.1}$$

La fonction indicatrice permet de définir le nombre d'occurrences $N_{\mathbf{X}}(\mathbf{w})$ d'un mot $\mathbf{w}$ dans $\mathbf{X}$:

$$N_{\mathbf{X}}(\mathbf{w}) = \sum_{i=1}^{l-k+1} Y_{\mathbf{X}}(i,\mathbf{w}) \tag{3.2}$$

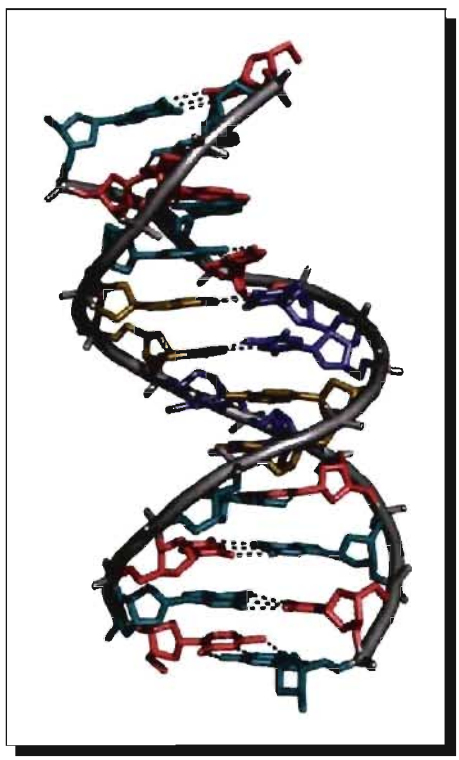
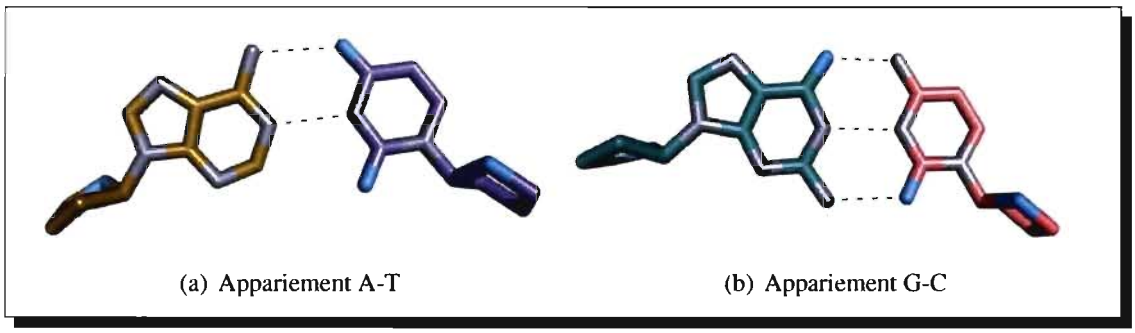


Figure 3.1 : Hélice d'ADN. L'adénine, la cytosine, la guanine et la thymine sont représentées en jaune, rose, vert et mauve, respectivement (image générée avec Pymol [20]).

3.2 Modèles statistiques

Notre technique de recherche d'éléments répétés se fonde sur l'idée élémentaire suivante : si une région est répétée plusieurs fois dans une séquence, la fréquence des oligonucléotides la composant sera anormalement élevée. Or, pour décider si la fréquence d'un mot $\mathbf{w}$ est anormalement élevée, on doit, au préalable, être en mesure de définir la fréquence d'oligonucléotides "normale" d'une séquence d'ADN aléatoire. Cette séquence aléatoire, appelée "modèle nul",



(a) Appariement A-T (b) Appariement G-C

Figure 3.2: Appariement canonique entre une adénine et une thymine (a) et entre une guanine et une cytosine (b). Les atomes de carbone de l'adénine, la cytosine, la guanine et la thymine sont représentés en jaune, rose, vert et mauve, respectivement, les atomes d'oxygène en bleu et les atomes d'azote en gris. Les liaisons d'hydrogènes sont représentées par un trait pointillé (images générées avec Pymol [20]).

est utilisée comme point de référence : si la fréquence d'un mot contenu dans une séquence observée est statistiquement supérieure à la fréquence du mot dans le "modèle nul ", on peut conclure que le mot est surreprésenté. De manière formelle, un mot $\mathbf{w}$ est statistiquement surreprésenté dans une séquence observée S si la probabilité $\mathbb{P}\{N_{\mathbf{X}}(\mathbf{w}) \geq N_{\mathbf{S}}(\mathbf{w})\}$ est inférieure à un seuil donné.

On retrouve plusieurs modèles de séquence aléatoire d'ADN dans la littérature. Cette section présente les principaux modèles utilisés.

3.2.1 Modèle de permutation

Le modèle de permutation est l'un des modèles les plus naturels pour construire une séquence aléatoire conservant la composition nucléique d'une séquence observée $\mathbf{S}$.

3.2.1.1 Permutation de nucléotides

Ce modèle considère l'ensemble $\mathcal{S}$ des permutations aléatoires de $\mathbf{S}$, c.-à-d. l'ensemble des séquences ayant la même longueur l que $\mathbf{S}$ et contenant le même nombre de a, c, g et t. Soit $\mathcal{I} = \{1, \ldots, l\}$, l'ensemble des indices de $\mathbf{S}$. Alors la fonction bijective

$$\rho : \mathcal{I} \mapsto \mathcal{I}$$

est appelée permutation. Une permutation aléatoire $\mathbf{X}$ de $\mathbf{S}$ est définie par

$$\mathbf{X} = (s_{\rho(i)})_{i \in \mathcal{I}}$$

3.2.1.2 Permutation d'oligonucléotides

Sachant que les protéines sont constituées d'acides aminés encodés par des triplets de nucléotides consécutifs, il est parfois avantageux d'effectuer les permutations sur de courts oligonucléotides plutôt que sur de simples nucléotides, de manière à capturer d'avantage d'information biologique potentielle. Soit t la longueur de l'oligonucléotide. Une permutation aléatoire $\mathbf{X}$ de taille t de $\mathbf{S}$ est définie par

$$\mathbf{X} = (s_{\rho^*(i)})_{i \in \mathcal{I}}$$

où ρ^* est restreint aux permutations préservant le compte des j-mers tel que $j \leq t$:

$$N_{\mathbf{X}}(\mathbf{w}) = N_{\mathbf{S}}(\mathbf{w}), \forall \mathbf{w} \in \mathcal{A}^{j=1\ldots t}$$

Une méthode utilisant le parcours eulérien d'un multigraphe dirigé permet de générer l'ensemble $\mathcal{S}$ [1, 33].

3.2.2 Modèle de Bernouilli

Une séquence d'ADN peut aussi être considérée comme une séquence finie de variables aléatoires indépendantes prenant une valeur $a \in \mathcal{A}$ avec une probabilité d'occurrence $\mu(a)$. Le modèle le plus simpliste associe des probabilités identiques aux différents caractères :

$$\mu(a) = \frac{1}{|\mathcal{A}|}, \forall a \in \mathcal{A} \tag{3.3}$$

alors que d'autres modèles utilisent la distribution des nucléotides dans une séquence de référence :

$$\mu(a) = \frac{N_{\mathbf{S}}(a)}{l}, \forall a \in \mathcal{A} \tag{3.4}$$

3.2.3 Modèle de Markov

Dans les chaînes de Markov, on considère encore une séquence

$$\mathbf{X} = X_1 \dots X_l$$

comme une suite de variables aléatoires prenant une valeur $a \in \mathcal{A}$. Contrairement au modèle de Bernouilli, les variables ne sont pas totalement indépendantes. La propriété de Markov statue que dans une chaîne de Markov d'ordre m, la variable X_i dépend uniquement des m dernières variables $(X_{i-m}, \dots, X_{i-1})$. Formellement, on a :

$$\begin{aligned} &\mathbb{P}\{X_i = a_i | X_1 = a_1, \dots, X_{i-m} = a_{i-m}, \dots, X_{i-1} = a_{i-1}\} \\ &\quad = \mathbb{P}\{X_i = a_i | X_{i-m} = a_{i-m}, \dots, X_{i-1} = a_{i-1}\} \end{aligned} \tag{3.5}$$

L'hypothèse d'homogénéité affirme que la séquence conserve les mêmes propriétés probabiliste du début à la fin, c.-à-d. la probabilité $\mathbb{P}\{X_i = a_i | X_{i-m} = a_{i-m}, \dots, X_{i-1} = a_{i-1}\}$ est indépendante de i.

Par convention, on dénote une chaîne de Markov d'ordre m par Mm. Le modèle de Bernouilli est en fait une chaîne de Markov d'ordre 0, ou M0, puisque les variables aléatoires sont indépendantes.

Une chaîne de markov d'ordre $m > 0$ est définie par une matrice carrée appelée matrice de transitions dénotée par $\Pi = (\pi(a_1 \dots a_m, a_{m+1}))_{a_1, \dots, a_{m+1} \in \mathcal{A}}$ tel que

$$\pi(a_1 \dots a_m, a_{m+1}) = \mathbb{P}\{X_i = a_{m+1} | X_{i-m} = a_1, \dots X_{i-1} = a_m\} \tag{3.6}$$

Un exemple de matrice de transitions d'une chaîne de Markov d'ordre 1 est affiché dans la Figure 3.3. La $i^{ème}$ rangée correspond à la distribution des probabilités de la variable X_{i+1} sous condition que la variable X_i égale i. Les valeurs de $\pi(a_1 \dots a_m, a_{m+1})$ satisfont les conditions

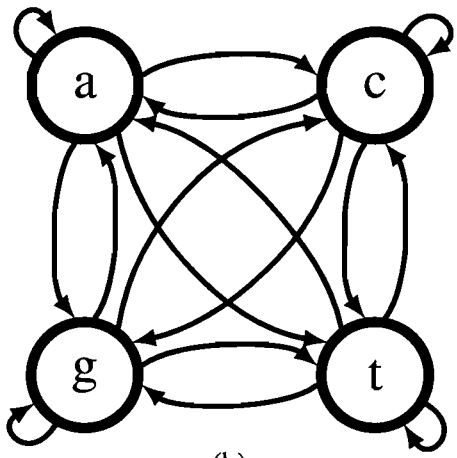
$$\pi(a_1 \dots a_m, a_{m+1}) \geq 0, \quad a_1, \dots, a_{m+1} \in \mathcal{A} \tag{3.7}$$

$$\sum_{a_{m+1} \in \mathcal{A}} \pi(a_1 \dots a_m, a_{m+1}) = 1 \tag{3.8}$$

L'équation (3.8) implique que la somme des probabilités d'une rangée de Π égale toujours 1.

$$\Pi = \begin{pmatrix} \pi(a,a) & \pi(a,c) & \pi(a,g) & \pi(a,t) \\ \pi(c,a) & \pi(c,c) & \pi(c,g) & \pi(c,t) \\ \pi(g,a) & \pi(g,c) & \pi(g,g) & \pi(g,t) \\ \pi(t,a) & \pi(t,c) & \pi(t,g) & \pi(t,t) \end{pmatrix}$$

(a)



(b)

Figure 3.3: Chaîne de Markov sur l'alphabet $\mathcal{A} = \{a,c,g,t\}$. (a) Matrice de transitions. (b) Représentation graphique des différentes transitions possibles.

Une fois le modèle M choisi, on désire estimer les paramètres $\theta = \{\theta_i\}$ en utilisant un ensemble de séquences d'entrainement $\mathcal{D}$, à condition que la chaîne de Markov soit stationnaire. L'estimation du maximum de vraisemblance est une méthode statistique qui permet de maximiser $P(\mathcal{D}|\theta,M)$ sur tous les θ possibles. De manière formelle, on écrit :

$$\theta^{\mathrm{MV}} = \arg\max_{\theta} P(\mathcal{D}|\theta,M)$$

Avec cette technique, il est possible de dériver les estimateurs des probabilités de transitions dénotées par $\hat{\pi}(a_1 \ldots a_m, a_{m+1})$:

$$\hat{\pi}(a_1 \ldots a_m, a_{m+1}) = \frac{N_{\mathrm{S}}(a_1 \ldots a_m a_{m+1})}{\sum_{b \in \mathcal{A}} N_{\mathrm{S}}(a_1 \ldots a_m b)} \tag{3.9}$$

Le désavantage de cette méthode est que lorsque l'ensemble d'entrainement est insuffisant, certains mots n'ont aucune occurrence, ce qui donne lieu à des probabilités de transition nulles. Par exemple, 24.94% des 4^{12} oligonucléotides de taille 12 n'ont aucune occurrence dans le chromosome 12 du génome humain. On peut partiellement pallier à ce problème en introduisant des données biologiques a priori à l'aide du théorème de Bayes. En supposant qu'il existe une distribution de probabilité sur les paramètres θ, on peut conditionner sur le modèle M pour

obtenir :

$$P(\theta|\mathcal{D},M) = \frac{P(\mathcal{D}|\theta,M)P(\theta|M)}{P(\mathcal{D}|M)}$$

où $P(\theta|\mathcal{D},M)$ et $P(\theta|M)$ dénote la probabilité a postériori et a priori, respectivement. On introduit l'estimateur du maximum a postériori (MAP) qui maximise la vraisemblance et la distribution a priori des paramètres θ :

$$\theta^{\text{MAP}} = \arg\max_{\theta} P(\mathcal{D}|\theta,M)P(\theta|M)$$

L'estimateur du maximum de vraisemblance est donc l'estimateur MAP pour une distribution a priori uniforme. En utilisant différentes distributions de probabilité a priori, on obtient différents estimateurs de probabilités. À titre d'exemple, en utilisant une distribution de Dirichlet avec paramètres $\alpha = (\alpha_a)_{a\in\mathcal{A}}$, on obtient :

$$\hat{\pi}(a_1 \ldots a_m, a_{m+1}) = \frac{N_{\mathbf{S}}(a_1 \ldots a_m a_{m+1}) + \alpha_{a_{m+1}}}{(\sum_{b\in\mathcal{A}} N_{\mathbf{S}}(a_1 \ldots a_m b)) + A} \tag{3.10}$$

où $A = \sum_{a\in\mathcal{A}} \alpha_a$. On retrouve d'autres exemples dans la littérature tel que l'utilisation de mixture de Dirichlet [11] ou de matrice de substitutions [15, 30].

Les deux modèles présentés conservent le compte de nucléotides. Cependant, le modèle de permutation conserve le compte exact alors que le modèle de Markov conserve l'espérance du compte. Comme nous le démontrons dans la section suivante, cette propriété n'est pas suffisante pour capturer la complexité des séquences biologiques lorsque l'on s'intéresse à la distribution des mots.

3.3 Distribution des occurrences de mots

Dans cette section nous sommes intéressés par la distribution statistique des occurences du mot $\mathbf{w}$ contenues dans une séquence aléatoire $X = (X_i)_{i=1,\ldots,l}$. Nous définissons la distribution de $N_{\mathbf{X}}(\mathbf{w})$ par

$$p_{\mathbf{w}}(n) = \mathbb{P}\{N_{\mathbf{X}}(\mathbf{w}) = n\}, \quad \forall n > 0$$

Dans une chaîne de Markov, la distribution exacte de $N_{\mathbf{X}}(\mathbf{w})$ peut être obtenue via la distribution de la position initiale T_n de la $n^{\text{ème}}$ occurence de w dans X. En effet, le nombre d'occu-

rences $N_{\mathbf{X}}(\mathbf{w})$ est plus grand ou égal à n si et seulement si la $i^{\text{ème}}$ occurence de w apparaît avant la $(l-k+1)^{\text{ème}}$ lettre de X :

$$N_{\mathbf{X}}(\mathbf{w}) \geq n \Longleftrightarrow T_n \leq l-k+1$$

et donc

$$\mathbb{P}\{N_{\mathbf{X}}(\mathbf{w}) = n\} = \mathbb{P}\{T_n \leq l-k+1\} - \mathbb{P}\{T_{n+1} \leq l-k+1\}$$

Robin *et al.* [67] utilisent une fonction récursive pour calculer la distribution de T_n et donc, par extension, la distribution exacte de $N_{\mathbf{X}}(\mathbf{w})$. Cependant, puisque cette technique nécessite un temps d'exécution considérable avec de longues séquences, des techniques d'approximation ont été développées. Lorsque l'espérance du nombre de mots $\mathbb{E}(N_{\mathbf{X}}(\mathbf{w}))$ est élevée, on utilise la distribution gaussienne. Si la distribution des lettres μ est connue, la moyenne $\mathbb{E}_{\mu}(N_{\mathbf{X}}(\mathbf{w}))$ et la variance $\mathbb{V}_{\mu}(N_{\mathbf{X}}(\mathbf{w}))$ peuvent être calculées (voir [38]). Puisque $N_{\mathbf{X}}(\mathbf{w})$ est en fait une somme de variables aléatoires (eq. 3.2), le théorème de la limite centrale permet d'établir la normalité asymptotique de la distribution de $N_{\mathbf{X}}(\mathbf{w})$. Lorsqu'au contraire, la distribution des lettres est inconnue et doit être approximée par $\hat{\mu}$, Prum *et al.* [65] proposent une autre approximation gaussienne qui tient compte de la dépendance entre le nombre d'occurrences observées et l'espérance du nombre d'occurrences $\mathbb{E}_{\hat{\mu}}(N_{\mathbf{X}}(\mathbf{w}))$. Pour les mots rares, la relation entre l'espérance du nombre d'occurrences et la longueur de la séquence n'est plus linéaire ; on approxime plutôt la distribution de $N_{\mathbf{X}}(\mathbf{w})$ par une distribution de Poisson composée [23, 69].

Plusieurs statistiques ont été développées pour comparer l'espérance théorique du nombre de mots $\mathbb{E}(N_{\mathbf{X}}(\mathbf{w}))$ au compte réel $N_{\mathbf{S}}(\mathbf{w})$ (voir [43] pour une revue de littérature). Une famille de statistiques souvent utilisée est appelée z-score :

$$z(w) = \frac{N_{\mathbf{S}}(\mathbf{w}) - \mathbb{E}(N_{\mathbf{X}}(\mathbf{w}))}{\sqrt{\mathbb{V}(N_{\mathbf{X}}(\mathbf{w}))}}$$

Le z-score est donc la normalisation de la différence entre le compte observé et le compte espéré du mot w ; une unité représente un écart type. Une valeur positive indique une surreprésentation alors qu'une valeur négative indique une sous-représentation du mot dans la séquence observée. L'utilisation de cette statistique permet de représenter les p-values sur une échelle gaussienne.

En pratique, le nombre d'occurrences d'un mot w dans un modèle aléatoire est souvent grandement sous-estimé ; plusieurs mots sont en réalité beaucoup plus fréquents dans les séquences biologiques que dans les séquences aléatoires générées avec les modèles précédemment présentés.

À titre d'exemple, la Figure 3.4 présente la distribution des fréquences de tous les oligonucléotides de taille 12, appelée "spectrum" dans [19]. L'axe des ordonnées affiche le nombre d'oligonucléotides, parmi les 4^{12} possibles, ayant une fréquence n, pour tous les n de l'axe des abscisses. Le spectrum de différentes séquences aléatoires est comparé à celui de la séquence biologique. Bien que la queue gauche de chacun des modèles aléatoires soit relativement bien ajustée à celle de la séquence biologique, on constate que la queue droite est très mal modélisée par les modèles simples ($m,t \leq 5$). Ceci est dû au fait que ces modèles supposent que le nombre d'oligonucléotides apparaissant le même nombre de fois décroit de manière exponentielle en fonction de la fréquence. On constate aussi qu'à ordre égal, la qualité de l'ajustement des chaînes de Markov est supérieure à celle des modèles de permutation. D'un point de vue technique, le modèle de permutation nécessite beaucoup plus de mémoire que le modèle de Markov (il faut plus de 8 gigaoctets de mémoire vive pour permuter le chromosome 22 avec t=10, alors que 2 gigaoctets son suffisants avec une chaîne de Markov d'ordre 10). Contrairement aux modèles d'ordre inférieur, la distribution de fréquence du modèle M10 est similaire à celle du chromosome 22. Cela s'explique cependant par le fait que l'ordre de la chaîne de Markov est similaire à la taille de l'oligonucléotide. Comme le démontre la Figure 3.5, la qualité de l'ajustement est inversement proportionnelle à la différence entre la taille de l'oligonucléotide et l'ordre de la chaîne. Dans la section suivante, une distribution capturant les caractéristiques intrinsèques des spectrums de divers organismes sera présentée.

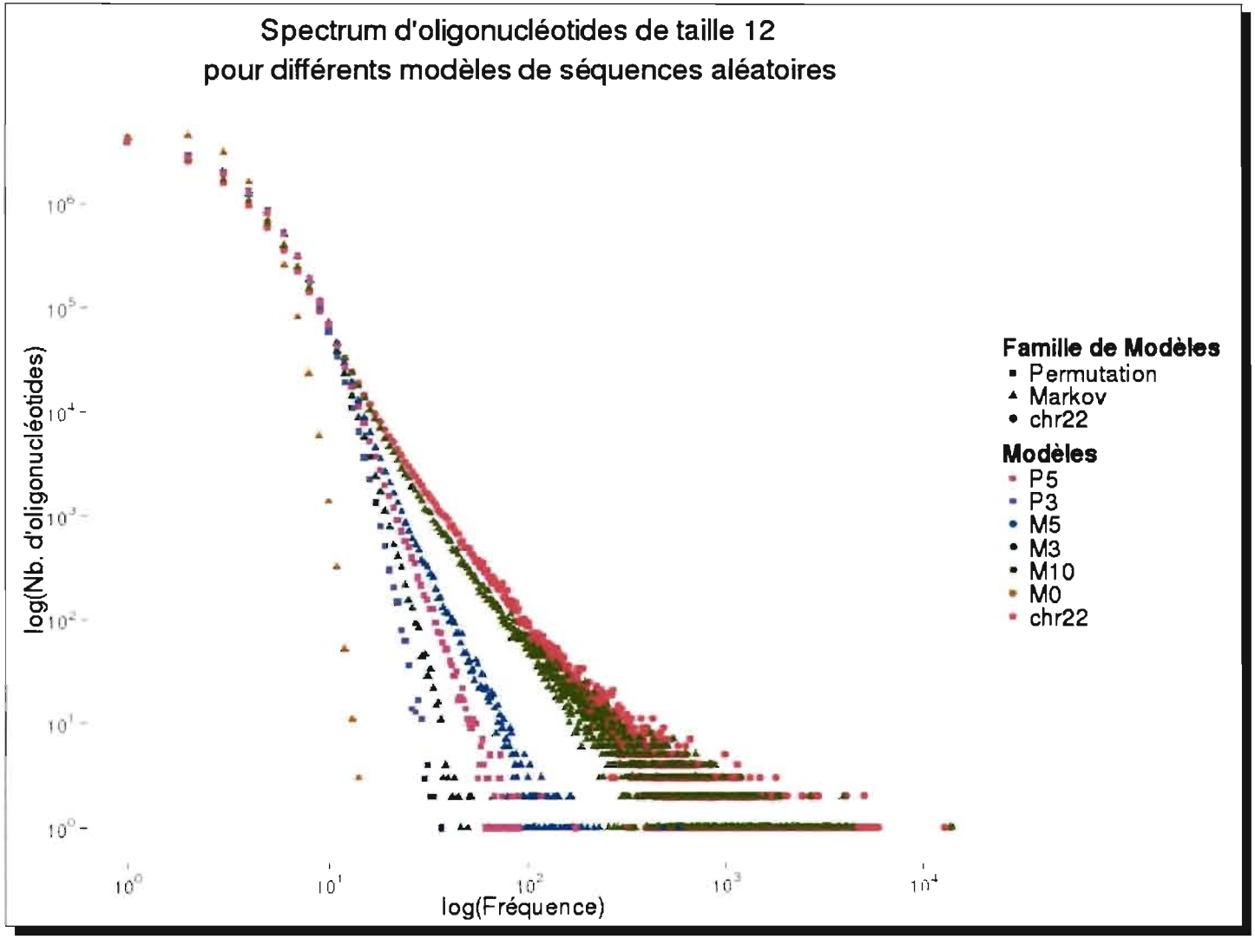


Figure 3.4: Spectrum d'oligonucléotides de taille 12 du chromosome 22 du génome humain. L'axe des abscisses représente le logarithme de la fréquence d'un oligonucléotide alors que l'axe des ordonnées représente le logarithme du nombre d'oligonucléotide ayant cette fréquence. La distribution des séquences générées par le modèle de permutation et par le modèle de Markov est affichée par les symboles ■, et ▲ respectivement alors que la distribution du chromosome 22 est affichée par le symbole •. Le modèle de permutation de bloc de taille k est dénoté par Pk alors que le modèle de Markov d'ordre m est dénoté par Mm.

3.4 Spectrum

L'approche présentée jusqu'à maintenant pour estimer la sur ou sous-représentation d'un mot consiste à comparer sa fréquence observée à sa fréquence espérée dans un modèle aléatoire de référence. Or, comme le démontre la Figure 3.4, les modèles de séquences aléatoires supposent que le nombre d'oligonucléotides apparaissant un même nombre de fois décroît exponentiellement avec le nombre d'occurrences. En réalité, la forme des spectrums des séquences biologiques suit une distribution log-normale alors que la queue droite suit une loi de puissance. Selon l'organisme considéré et la taille du mot, la queue gauche peut aussi suivre une loi de puissance. De fait, l'utilisation de séquences aléatoires comme modèle nul n'est pas appropriée pour générer la distribution des occurrences de mots. La technique qui sera utilisée dans cette étude consiste plutôt à utiliser une distribution d'occurrences de mots comme modèle de référence.

3.4.1 Loi de puissance

Historiquement, on a observé que la fréquence f d'un mot faisant partie du corpus d'une langue naturelle est inversement proportionelle à son rang R dans la table des fréquences; $f = aR^{-k}$ avec $k \sim 1$ [87]. Ce comportement, observé dans plusieurs distributions de population, fait partie de la famille des distributions appelées "loi de puissance" (de l'anglais *Power Law*).

Mantegna *et al.* [52] ont affirmé, en traçant la fréquence des l-mers (mots de longueur l) en fonction de leur rang, pour $3 \leq l \leq 8$, que la queue de la distribution des occurences de mots d'ADN suivait aussi une loi de Zipf. Ils ont de plus affirmé que l'ADN non codant avait plus de ressemblances avec les langues naturelles que les régions codantes et, par extension, que l'ADN non codant peut possiblement contenir un ou des languages biologiques structurés. Cette dernière affirmation a soulevé de nombreuses critiques. On a entre autres affirmé que le fait que l'ADN suive une loi de Zipf n'implique pas un language sous-jacent puisque plusieurs types de distribution suivent une telle loi (taille des villes, salaires, nombre de publications par auteur) [41], qu'une plus grande variance dans la fréquence des nucléotides, dans les régions non codantes, explique les observations [7, 8] et que le test linguistique ne permet pas de différencier quantitativement une séquence d'ADN d'une séquence artificielle [14]. Martindale *et*

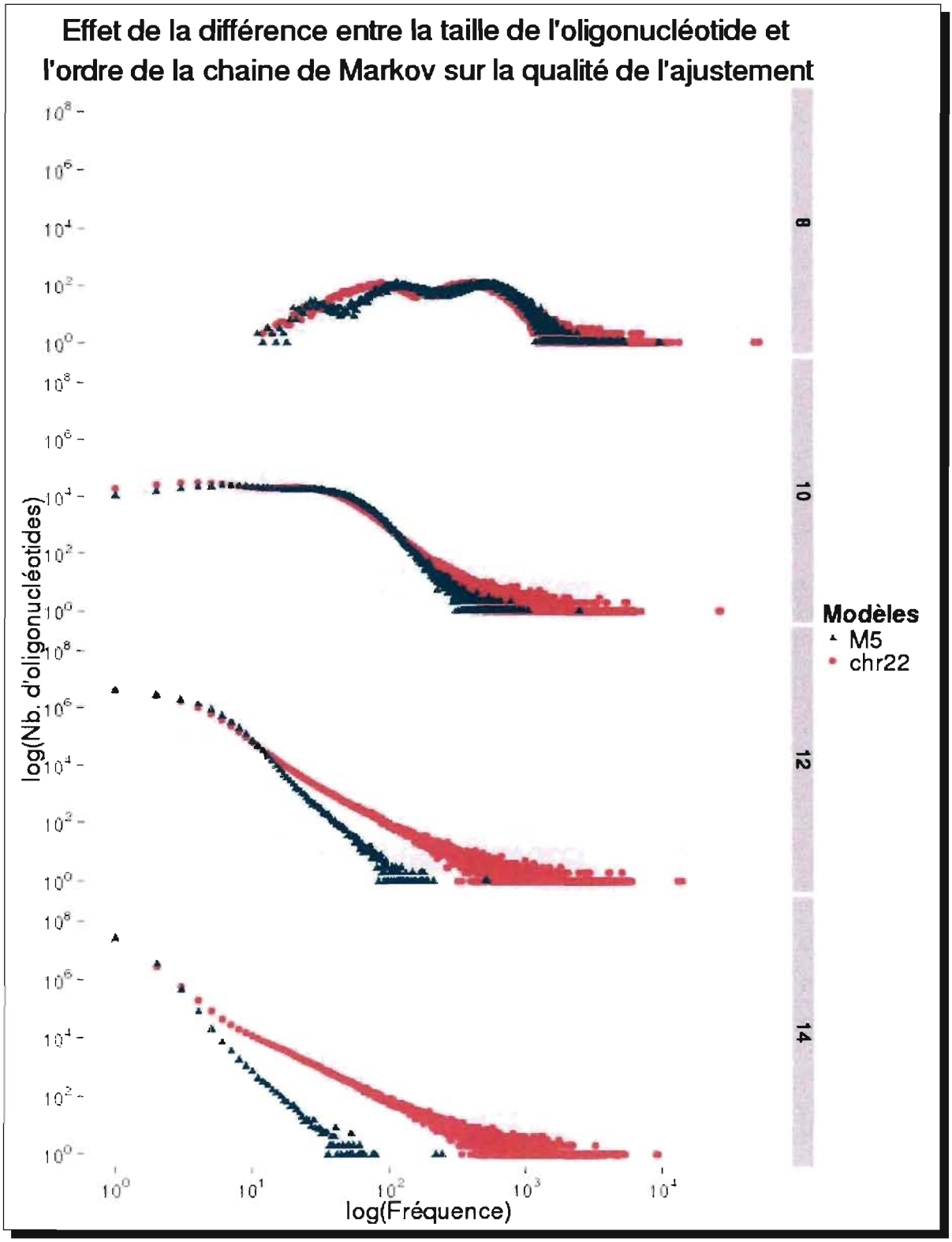


Figure 3.5: La différence entre la taille de l'oligonucléotide et l'ordre de la chaîne de Markov est inversement proportionnelle à la qualité de l'ajustement. Distribution de fréquence d'oligonucléotide de taille 8, 10, 12 et 14 pour une séquence aléatoire générée avec une chaîne de Markov d'ordre 5 et pour le chromosome 22 du génome humain.

al. [53] ont, quant à eux, décelé des erreurs de méthodologie. Dans [52], les fréquences suivent une loi de puissance pour les 1000 premiers rangs. Or, lorsque la majorité des points sont situés dans la queue d'un échantillon, il est impératif de choisir une distribution qui s'ajuste sur l'intervalle complet plutôt que sur un sous-intervalle. On affirme donc dans [53] que la queue des fréquences ordonnées d'oligonucléotides suit plutôt une distribution de Yule définie par l'équation :

$$f = aR^{-k}b^{R}$$

Ils affirment de plus que l'ajustement est adéquat pour des oligonucléotides de taille l avec $5 \leq l \leq 9$. Cependant, cette étude porte sur de très courtes séquences ; les séquences procaryotiques, levuriformes, invertébrées et humaines totalisent 784344, 315338, 120966 et 824025 nucléotides respectivement. Ces résultats ne sont donc pas valides pour l'ordre de grandeur qui nous intéresse dans cette étude.

Dans [19] et [48], on s'intéresse plutôt à la distribution des fréquences d'oligonucléotides de même longueur l. Luscombe *et al.* [48] ont, en partie, confirmé les résultats de [53] en démontrant que la queue droite du spectrum de l-mer, avec $6 \leq l \leq 10$, suit un loi de puissance. Ils ont, de plus, constaté que la pente de de la queue est similaire pour toute taille l ; le nombre de l-mers ayant une fréquence donnée décroît à un rythme similaire, indépedemment de la taille l. Ceci peut s'expliquer par le fait que les distributions ne sont pas indépendantes : les mots plus longs contiennent une combinaison des mots plus courts. On explique aussi que les l-mers avec $l < 6$ ne suivent pas une loi de puissance puisqu'il n'y a pas assez de mots distincts pour différencier les niveaux d'occurrences.

3.4.2 Distribution double Pareto log-normale

Dans un article plus récent, Csűrös *et al.* [19] ont démontré que le spectrum des oligonucléotides suivait une distribution double Pareto log-normale (DPLN) [66]. Cette distribution, contrairement à la distribution de Yule ou à la loi de puissance qui ne modélise que la queue de la distribution, modélise la totalité du spectrum. Comme son nom l'indique, la distribution DPLN est constituée d'une région centrale suivant une distribution log-normale avec

paramètres ν et τ borné par une queue gauche et droite suivant une distribution Pareto avec paramètres β et α respectivement. Soit D, une variable aléatoire suivant la distribution double Pareto log-normale :

$$D \sim \text{dpln}(\alpha, \beta, \nu, \tau^2)$$

Cette variable aléatoire peut être représentée par

$$D = \frac{X_1}{X_2} Y$$

où X_1, X_2 et Y sont des variables aléatoires indépendantes tel que :

- $\log Y \sim \text{N}(\nu, \tau^2)$
- $X_1 \sim \text{Pareto}(\alpha)$
- $X_2 \sim \text{Pareto}(\beta)$

On peut alternativement décrire la variable D par le produit

$$D = YR$$

où R est le ratio des distribution Pareto $\frac{X_1}{X_2}$. Puisque la distribution Pareto de paramètre θ possède une fonction de densité de probabilité définie par

$$f(\nu) = \theta \nu^{-\theta-1},$$

la fonction de densité de probabilité de la variable R est :

$$f(r) = \begin{cases} \frac{\alpha\beta}{\alpha+\beta} r^{b-1} & \text{pour } 0 < r \leq 1 \\ \frac{\alpha\beta}{\alpha+\beta} r^{-a-1} & \text{pour } r > 1 \end{cases} \tag{3.11}$$

Donc, lorsqu'affichée avec des échelles logarithmiques, les pentes des queues gauche et droite sont données par les paramètres $\beta - 1$ et $-\alpha - 1$ respectivement (Figure 3.6). L'influence des différents paramètres sur la forme de la distribution est affichée dans la Figure 3.7.

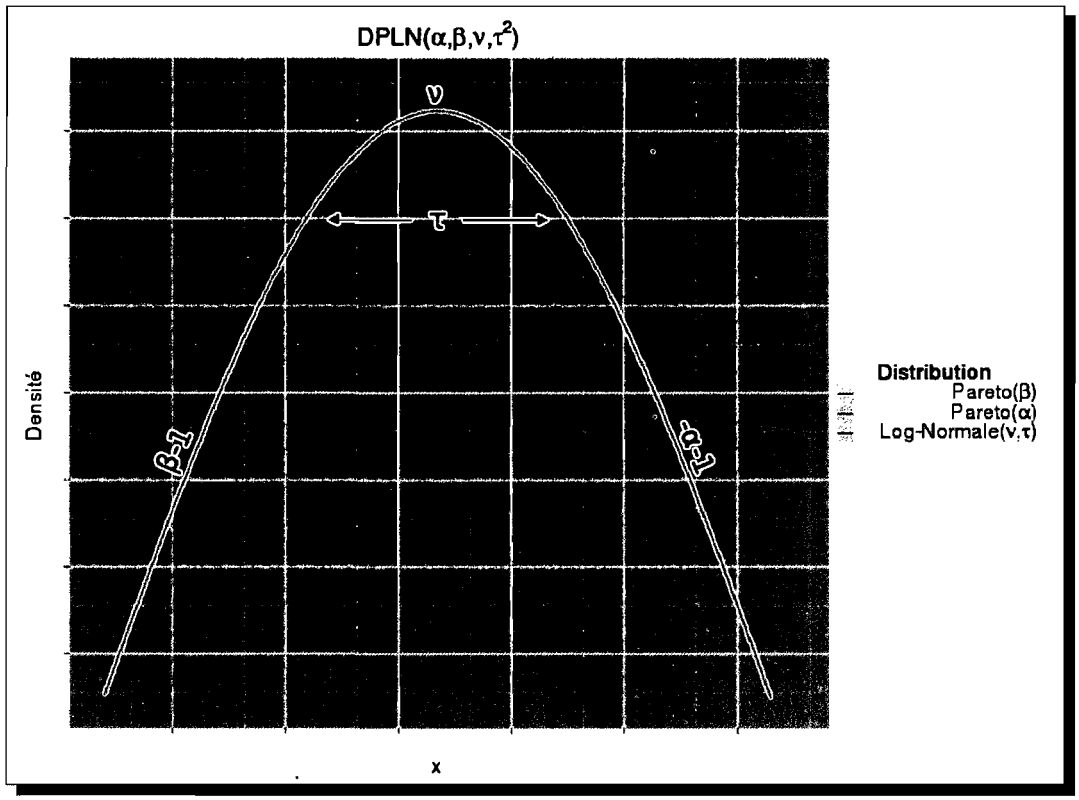


Figure 3.6: Distribution double Pareto log-normale de paramètres $\alpha = 0.5, \beta = 2.5, \nu = 6$ et $\tau = 1$. La distribution DPLN est constituée d'une distribution log-normale de paramètres ν et τ bornée à gauche et à droite par deux distributions Pareto de paramètre β et α respectivement. Lorsqu'affichée avec des axes logarithmiques, la pente de la queue gauche égale $\beta - 1$ alors que celle de la queue droite égale $-\alpha - 1$. Le paramètre τ influe sur l'évasure de la distribution, c.-à-d. la largeur de la "cloche", alors que le paramètre ν influe sur la position de la distribution.

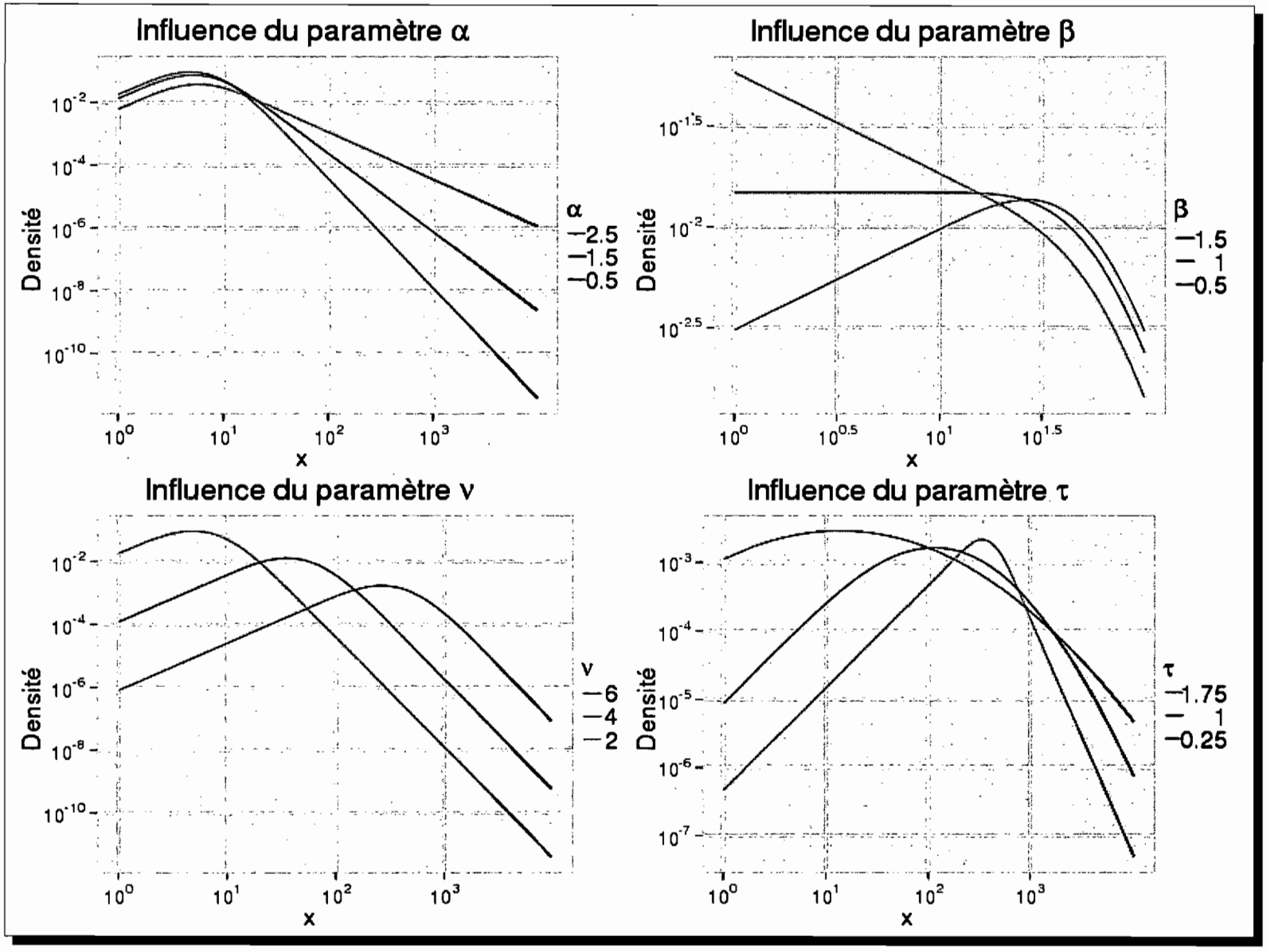


Figure 3.7: Influence des paramètres α, β, ν, τ sur la forme de la distribution double Pareto log-normale.

3.4.3 Spectrum pour différentes tailles de mots et de séquences

Cette distribution composite permet de modéliser le spectrum des séquences biologiques de divers organismes, pour un grand intervalle de tailles de k-mer et de séquences. La Figure 3.9 affiche le spectrum du chromosome 12 du génome humain pour différentes tailles d'oligonucléotides alors que le spectrum de différents chromosomes humains pour un oligonucléotide de taille 12 est affiché dans la Figure 3.10. Afin d'inspecter l'influence de la taille des mots et des séquences sur les paramètres de la distribution, ceux-ci ont aussi été affichés dans la Figure 3.11. Finalement, les paramètres de la régression linéaire effectuée dans la Figure 3.11 sont affichés dans la table 3.I.

3.4.3.1 Influence de la taille des mots

Les régressions linéaires sur les paramètres α et τ (Fig.3.11 haut, table 3.I gauche) ne sont pas statistiquement significatives avec des R^2 respectifs de 0.221 et 0.251. Cependant, les régressions linéaires des paramètres β et ν sont très significatives ($R^2 = 0.993$ pour les deux régressions linéaires), impliquant une relation linéaire inversement proportionnelle entre la taille de l'oligonucléotide et le paramètre β d'une part, et entre la taille de l'oligonucléotide et le paramètre ν d'autre part. Ces paramètres sont dépendants ; la fréquence moyenne (ν) d'un mot augmente lorsque sa taille diminue, ce qui a pour effet de déplacer la distribution vers la droite. Lorsque la distribution se déplace vers la droite, la pente de la queue gauche (β) augmente, puisque moins de mots ont une faible fréquence. La Figure 3.8 démontre la dépendance de ν et β en affichant une régression linéaire significative ($R^2 = .986$).

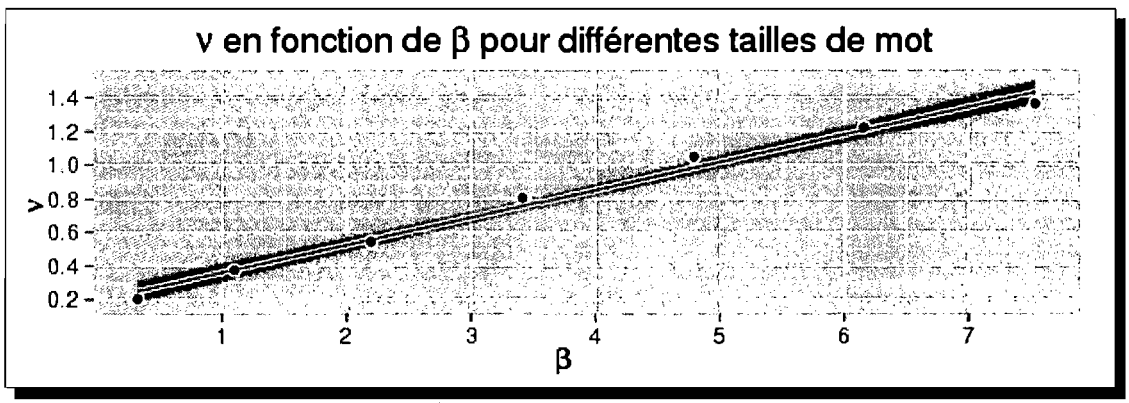


Figure 3.8: Régression linéaire entre les paramètres β et ν pour différentes tailles de mots.

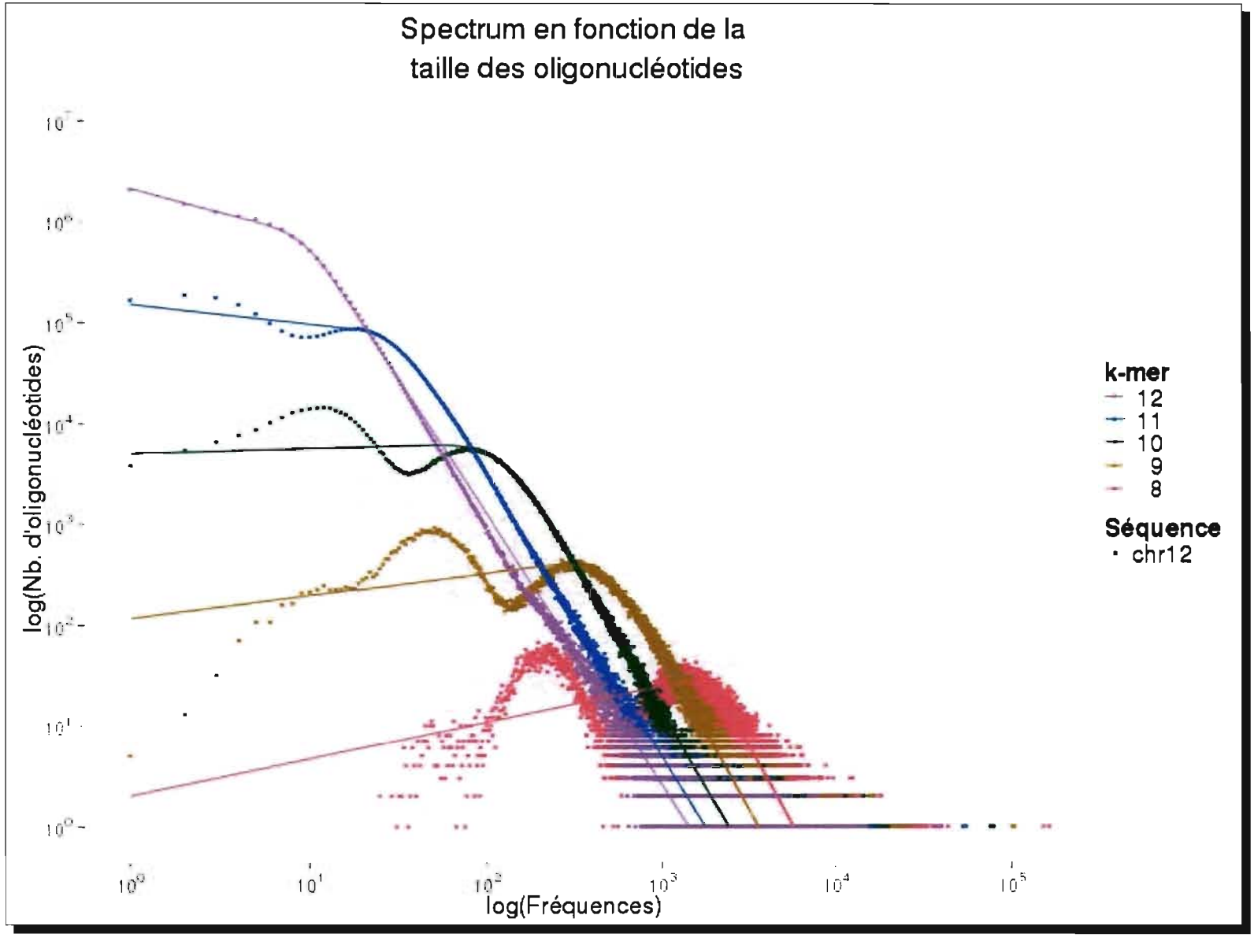


Figure 3.9: Distribution double Pareto Log-Normale ajustée au spectrum du chromosome 12 du génome humain pour différentes tailles d'oligonucléotides.

3.4.3.2 Influence de la taille des séquences

Les régressions linéaires (Figure 3.11 bas, table 3.I droite) ne sont pas significatives étant donné la différence d'ordre de grandeur entre la valeur des paramètres et la longueur des séquences. Cependant, une inspection visuelle permet de constater une augmentation linéaire certaine du paramètre ν lorsque la taille de la séquence augmente. Les trois autres paramètres varient très peu ; la taille de la séquence influe principalement sur un seul paramètre. Dans la section 5.1.3.4, nous expliquerons comment notre technique de recherche tire profit de cette particularité.

Paramètre	OO	Pente	R^2
α	2.390	-0.040	0.221
β	3.020	-0.202	0.993
ν	17.150	-1.228	0.993
τ	0.182	0.012	0.251

Paramètre	OO	Pente	R^2
α	1.717	1.1e-10	1.55e-02
β	0.447	6.2e-10	3.84e-01
ν	0.919	8.5e-09	8.84e-01
τ	0.269	9.1e-12	8.18e-05

Tableau 3.I: Paramètres des régressions linéaires effectuées dans le graphique 3.11. Les paramètres des régressions effectuées sur α β ν et τ des distributions DPLN sur la taille des mots (Figure 3.11 haut) et sur la taille des séquences (Figure 3.11 bas) sont affichés à gauche et à droite, respectivement.

La dernière partie de ce chapitre traite des graines espacées, technique par laquelle il est possible d'augmenter la sensibilité d'une recherche sans en diminuer la spécificité.

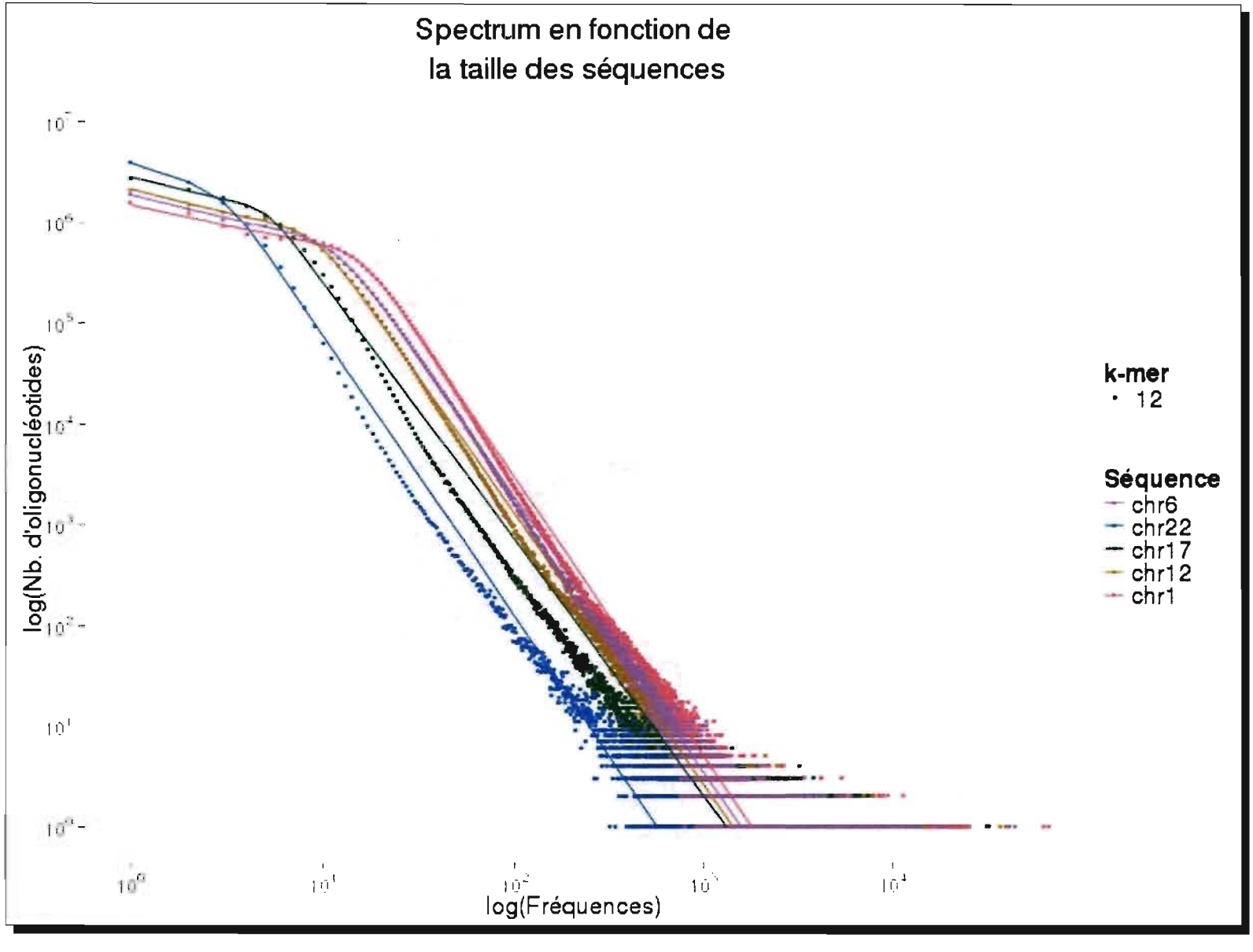


Figure 3.10: Distribution double Pareto Log-Normale ajustée aux spectrums des chromosomes 1, 6, 12, 17, et 22 du génome humain pour des oligonucléotides de taille 12.

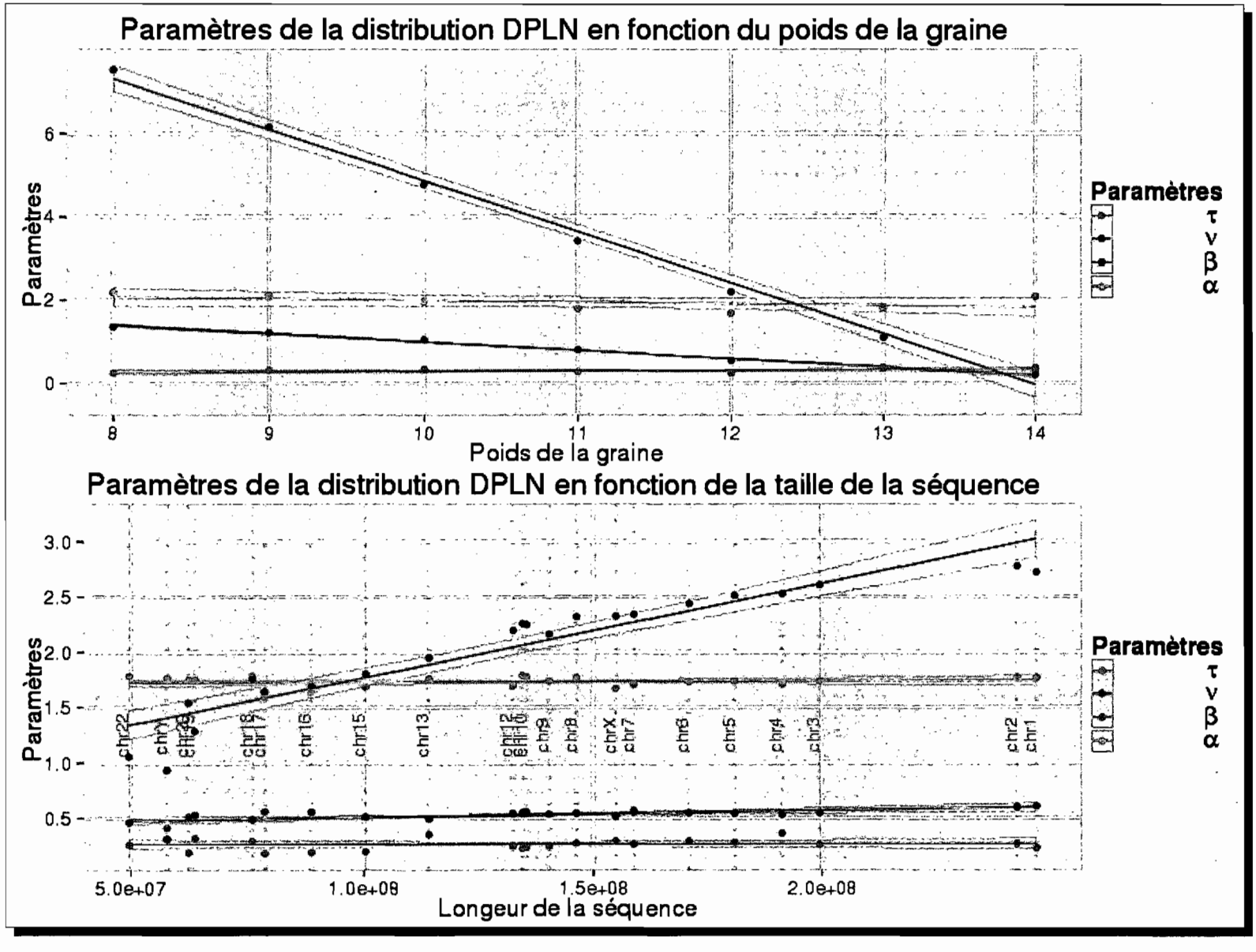


Figure 3.11: Paramètres des distributions DPLN en fonction de la taille du mot de la Figure 3.9 (haut) et de la taille de la séquence de la Figure 3.10(bas). Les paramètres de la régression linéaire sont affichés dans la Table3.I

3.5 Graines espacées

La recherche d'homologie dans les applications traditionnelles telle que BLAST [2] s'effectue en deux étapes. En premier lieu, on recherche des graines, c.-à-d. des régions identiques, contenues dans la séquence et dans la requête. En second lieu, les graines trouvées lors de la première étape sont étendues pour trouver des régions homologues.

Dans leur application PatternHunter, Ma *et al.* [50], ont définis le concept de graines espacées. Une graine espacée g est un ensemble de 1 et 0, où un 1 implique que les lettres de la séquence cible et de la requête doivent correspondre, alors qu'un 0 indique que les lettres peuvent ou non correspondre. Une graine espacée débute et se termine par un 1. On définit le poids par le nombre de 1 alors que la longueur est définie par le nombre de 1 et de 0. Par exemple, la graine $g = 100111$ possède un poids de 4 et une taille de 6. En utilisant cette graine, les séquences AAAAAA et ACCAAA s'alignent puisque les caractères aux positions définies par la graine espacée g sont identiques. Il a été démontré dans [50] que l'utilisation de graines espacées permet d'augmenter la sensibilité d'une recherche, i.e. d'augmenter la probabilité d'avoir au moins un "*hit*" dans une région donnée tout en diminuant l'espérance du nombre de "*hits*" aléatoires, par rapport aux graines consécutives. On peut intuitivement expliquer ce gain en sensibilité par le fait que les "hits" consécutifs sont plus probables puisque plus indépendants lorsque l'on utilise une graine espacée. En effet, le nombre de bases ré-échantillonnées entre une graine débutant en position i et la même graine débutant en position $i+1$ est plus petit lorsqu'on utilise une graine espacée que lorsque l'on utilise une graine consécutive (Figure 3.12).

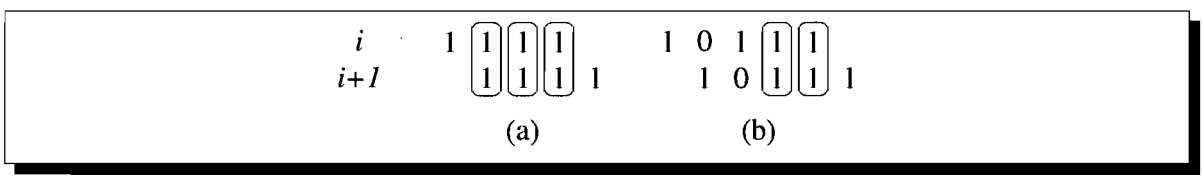


Figure 3.12: Le nombre de positions ré-échantillonnées entre une graine débutant en position i et la même graine débutant en position $i+1$ est plus petit lorsque l'on utilise une graines espacées (b) que lorsque l'on utilise une graine consécutive (a) de même poids. Un nombre de position ré-échantillonné plus petit implique une plus grande indépendance et, par conséquent, une plus grande sensibilité.

Buhler *et al.* [12, 75], ont développé une application appelée Mandala qui permet de générer un ensemble de graines espacées de manière à maximiser la sensibilité mutuelle ; c.-à-d. une fois la première graine sélectionnée, l'algorithme recherche une seconde graine qui maximise la sensibilité commune, et ainsi de suite. À l'aide de cet algorithme, nous avons généré des ensembles de 10 graines espacées pour différents poids. L'ensemble de graines espacées de poids 12 utilisées dans notre étude est le suivant :

Id	Graine espacée
A	1110101001011001111
B	1111001001100010100111
C	1101011101000111011
D	1101011010011000010111
E	1110110000101010011011
F	1101000110010110101011
G	1111001110111101
H	1110010100000110110111
I	1101100101111001000111
J	1101010100101001101011

Tableau 3.II: L'ensemble de graines espacées de poids 12.

En calculant le spectrum d'une séquence avec plusieurs graines espacées de même poids mais de longueurs et configurations distinctes, on espère obtenir un spectrum consensus qui est plus sensible et de fait, moins vulnérable aux artéfacts biologiques.

Notre première hypothèse est qu'il existe une corrélation entre la distribution des occurrences de mots produites en utilisant deux graines espacées de même poids, mais de conformation et de longueur distinctes. Pour ce faire, on parcourt la séquence d'ADN en calculant, pour chaque position, la fréquence globale du mot extrait à cette position avec une graine espacée. On calcule ensuite le coefficient de corrélation de Pearson entre les séquences de fréquences produites par chaque paire de graines espacées distinctes. La Figure 3.13 affiche les résultats obtenus pour les graines de la Table 3.II. Tous les coefficients prennent une valeur telle que $0.4 \lesssim \rho \lesssim 0.7$. Les séquences de fréquences produites par les graines espacées G, A et C sont, dans l'ordre, celles qui sont le moins corrélées aux autres (ρ moyens de 0.498, 0.587 et 0.595, respectivement). Ceci s'explique probablement par le fait que ces trois graines sont plus courtes que les 7 autres. La moyenne globale des 45 coefficients de corrélation de Pearson est

de 0.604, ce qui implique une bonne corrélation entre les différentes séquences de fréquences. À l'opposé, le fait que les coefficients ne soient pas trop élevés implique aussi une certaine indépendance entres les graines, ce qui laisse présager que l'utilisation de plusieurs graines distinctes permette d'améliorer la sensibilité de la technique de recherche.

La deuxième hypothèse suppose que les spectrums produits par différentes graines espacées de même poids sont similaires. Dans la Figure 3.14, on compare le spectrum du chromosome 12 produit par les graines espacées G-H, G-E, G-B, G-D ; les 4 paires de graines ayant obtenu les plus faibles coefficients de corrélation (0.414, 0.428, 0.453 et 0.485, respectivement). Malgré le fait que les graines soient très différentes, les spectrums sont très similaires. Dans le même ordre d'idée, la Figure 3.15 affiche la distribution DPLN produite par l'ensemble complet des 10 graines espacées alors que la Figure 3.16 affiche les paramètres de ces 10 distributions.

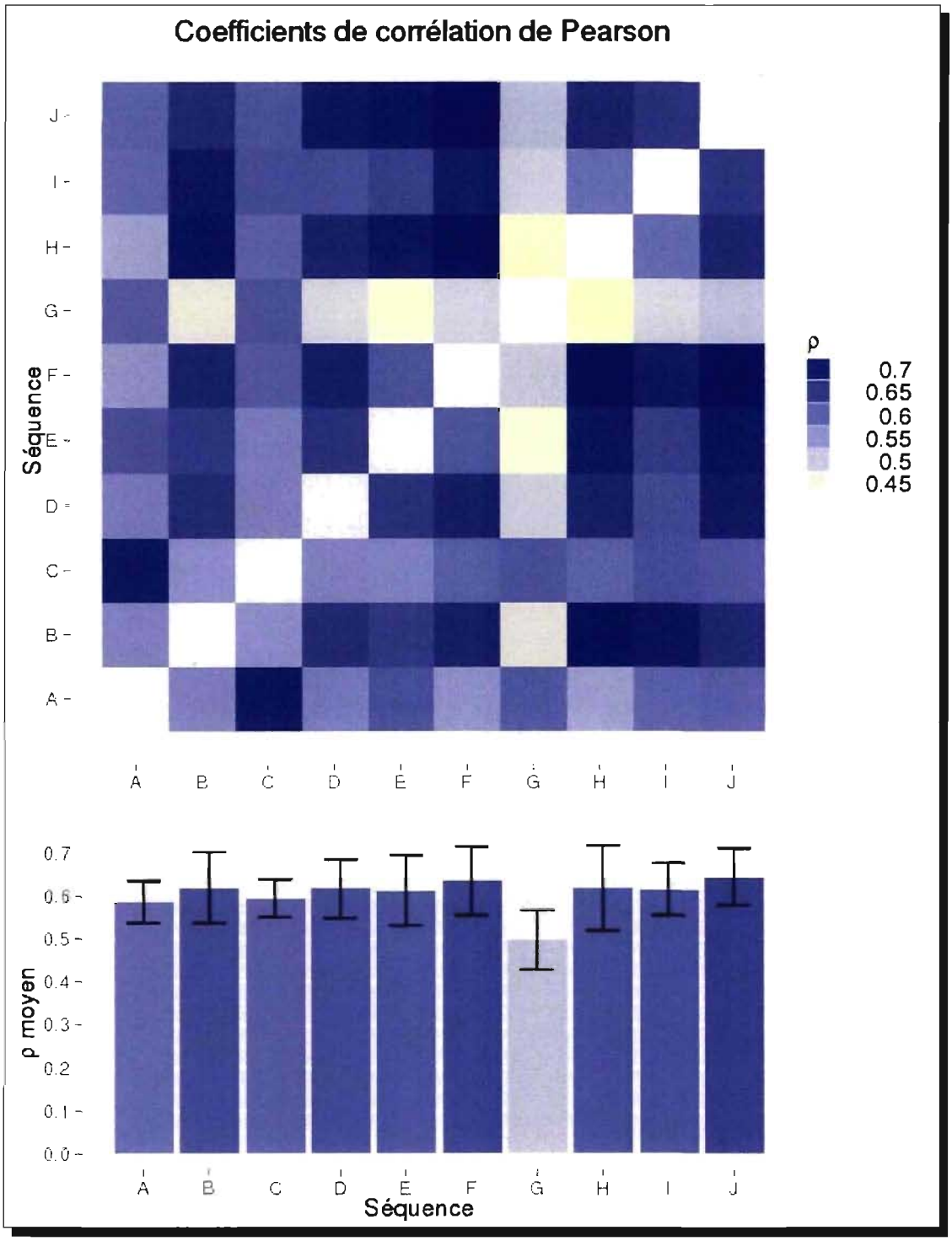


Figure 3.13: Comparaison binaire des coefficients de Pearson des spectrums du chromosome 12 générés par 10 graines espacées distinctes.

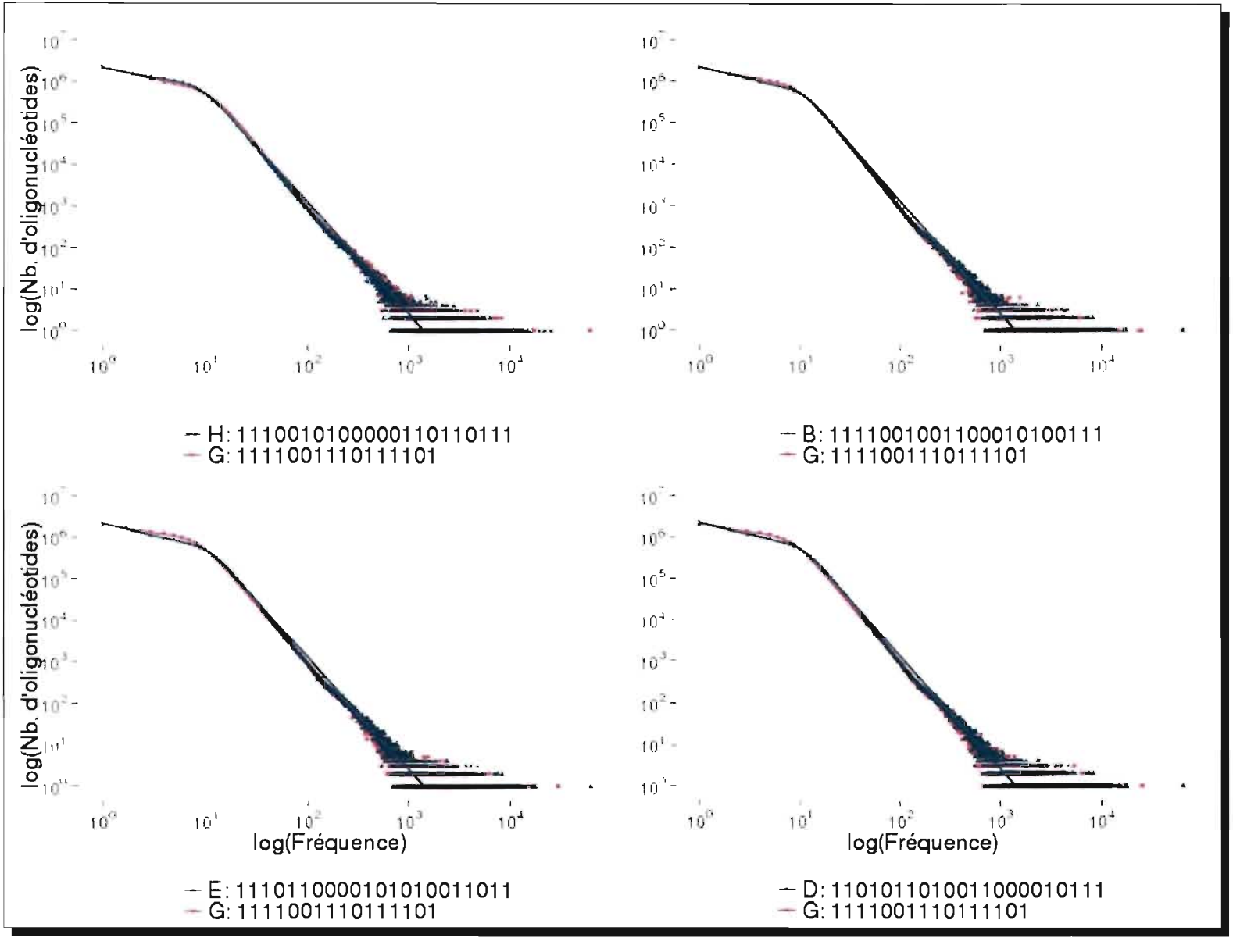


Figure 3.14: Comparaison de certains spectrums du chromosome 12 générés par différentes graines espacées.

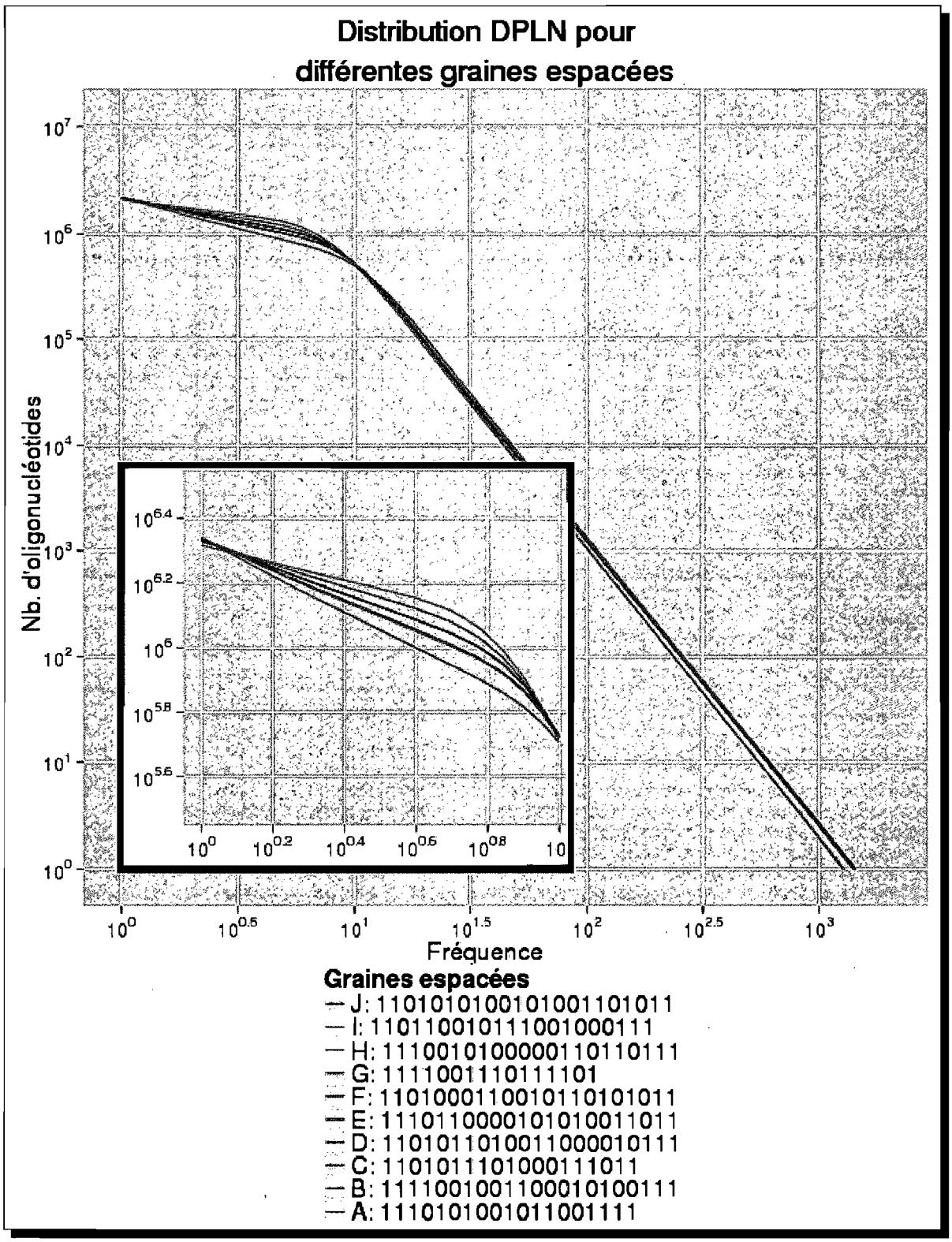


Figure 3.15: Distribution DPLN des spectrums du chromosome 12 générés par 10 graines espacées distinctes.

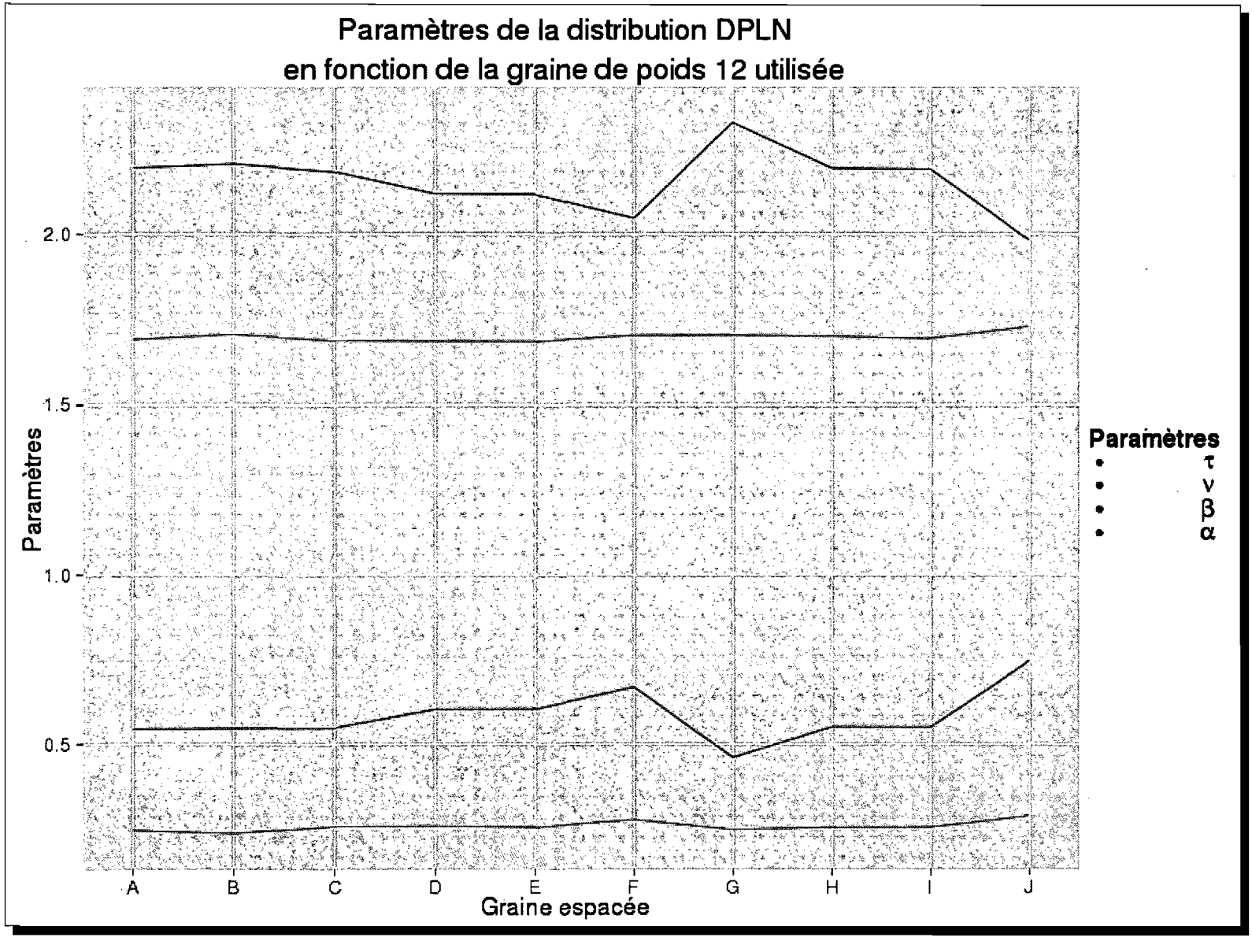


Figure 3.16: Paramètres des distribution DPLN de la Figure 3.15.

CHAPITRE 4

IDÉE PRINCIPALE ET CONCEPT CLÉ

Dans cette section, nous définissons le problème de recherche de régions répétées dans une séquence d'ADN et présentons l'idée principale de notre solution proposée.

4.1 Définition du problème

Étant donné $\mathbf{S} = s_1 ... s_n$ où $s_i \in \mathcal{A} = \{a, c, g, t\}$, une séquence d'ADN cible et $\mathbf{T} = t_1 ... t_m$ où $t_i \in \mathcal{A} = \{a, c, g, t\}$, une séquence d'ADN d'entraînement. On définit par $\mathcal{I}_S = \{1, \ldots, n\}$ et $\mathcal{I}_T = \{1, \ldots, m\}$ les ensembles d'indice des séquences **S** et **T** respectivement. On possède un échantillon d'entraînement $H_T : \mathcal{I}_T \rightarrow \mathcal{Y}$ où $\mathcal{Y} = \{\text{Répété, Non Répété}\}$, qui permet d'annoter la séquence d'entraînement **T**. Le problème consiste à trouver un classificateur $H_S : \mathcal{I}_S \rightarrow \mathcal{Y}$ qui permet d'associer chacun des indices de $\mathcal{I}_S$ à une étiquette $y \in \mathcal{Y}$ de manière à maximiser la vraisemblance étant donné **T** et son classificateur H_T.

4.2 Idée principale

L'idée principale consiste à utiliser un modèle de Markov caché pour segmenter la séquence observée en deux classes *R* et *NR* correspondant aux régions répétées et non répétées respectivement. Au lieu d'utiliser la séquence **S** comme processus observé, le modèle de Markov caché utilise la séquence de fréquences d'oligonucléotides ; on associe à chaque position la fréquence totale de l'oligonucléotide débutant à cette position (Figure 4.1). Les probabilités d'émission sont estimées grâce à un modèle nul de distribution d'oligonucléotides, à savoir, la distribution double Pareto log-normale. On peut ainsi inférer l'annotation des régions répétées en utilisant les probabilités d'état à postériori, c.-à-d. la probabilité qu'une fréquence f_i appartienne à une région donnée étant donné la séquence observée et le modèle utilisé.

4.3 Définition des concepts clés

Nous définissons et redéfinissons ici quelques concepts clés pour introduire notre solution proposée.

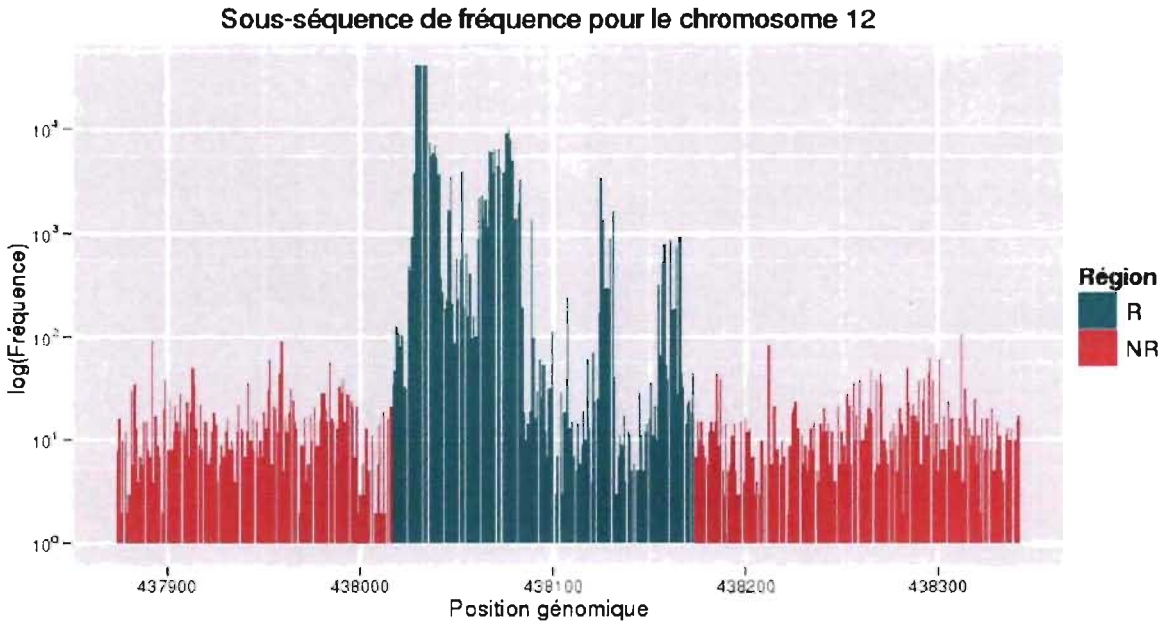


Figure 4.1: Séquence de fréquences d'oligonucléotides d'une sous-séquence du chromosome 12. L'axe des x représente la position chromosomique, l'axe des y, le logarithme de la fréquence de l'oligonucléotide débutant à cette position. On constate que la fréquence des oligonucléotides est beaucoup plus élevée dans les régions répétées (vert) que dans les régions non répétées (rose). L'algorithme tire partie de cette différence de distribution pour segmenter la séquence.

Définition 1. *Une* séquence *d'ADN est représentée par* $\mathbf{S} = s_1 \dots s_n$, *où n correspond à la longueur de la séquence et où* $s_i \in \mathcal{A} = \{a, c, g, t\}$.

Définition 2. *Un* mot *contenu dans une séquence est représenté par* $\mathbf{w} = w_1 \dots w_k$, *où k est la taille de* $\mathbf{w}$ *et où* $w_i \in \mathcal{A} = \{a, c, g, t\}$. *On dénotera par* $\mathbf{w}_\mathbf{S}(i)$ *le mot de taille k débutant à la position i dans la séquence* $\mathbf{S}$, *c.-à-d.*

$$\mathbf{w}_\mathbf{S}(i) = s_i \dots s_{i+k-1}$$

Par exemple si on a une séquence $\mathbf{S} = aacgtagattt$ alors $\mathbf{w}_\mathbf{S}(4) = gtaga$ pour $k = 5$.

Définition 3. *Une* graine espacée *est définie comme dans [76] par une liste d'indices* $\mathbf{g} = \{g_1, \dots, g_p\}$ *avec* $g_1 = 0$. *Le nombre de positions inspectées p définit le* poids *de la graine espacée alors que sa* longueur *l égale* $g_p + 1$. *Pour toute position j tel que* $0 \leq j \leq l$, *si* $j \in \mathbf{g}$ *alors le caractère à cette position est échantillonné ; autrement, le caractère est exclu. Cette*

définition est préférée à celle équivalente du vecteur caractéristique proposée à la section 3.5 puisqu'elle permet de simplifier les définitions ultérieures.

Par exemple la graine espacée $\mathbf{g} = \{0,1,4,6,7\}$ possède un poids de 5 et une longueur de 8.

Définition 4. *Étant donné une graine espacée* $\mathbf{g}$, *un* mot espacé *est définit par* $\mathbf{w}^{\mathbf{g}} = w_1 \dots w_p$, *où p est la taille de* $\mathbf{w}^{\mathbf{g}}$ *(par définition, p est le poids de la graine* $\mathbf{g}$*) et où* $w_i \in \mathcal{A} = \{a,c,g,t\}$. *On dénotera par* $\mathbf{w}^{\mathbf{g}}_{\mathbf{S}}(i)$ *le mot espacé extrait à la position i de la séquence* $\mathbf{S}$ *avec la graine g, c.-à-d.*

$$\mathbf{w}^{\mathbf{g}}_{\mathbf{S}}(i) = s_i s_{i+g_2} \dots s_{i+g_p}$$

Par exemple, en utilisant la séquence $\mathbf{S} = aacgtagattt$ et la graine espacée définie plus haut, $\mathbf{w}^{\mathbf{g}}_{\mathbf{S}}(3) = \{cggtt\}$

Définition 5. *La fonction indicatrice d'un mot* $\mathbf{w}$ *situé à la position i dans* $\mathbf{S}$ *est dénotée par*

$$Y_{\mathbf{S}}(i,\mathbf{w}) = \begin{cases} 1 & si\ s_i \dots s_{i+k-1} = w_1 \dots w_k \\ 0 & sinon \end{cases} \tag{4.1}$$

Similairement, la fonction indicatrice d'un mot espacé $\mathbf{w}^{\mathbf{g}}$ *est donnée par*

$$Y_{\mathbf{S}}(i,\mathbf{w}^{\mathbf{g}}) = \begin{cases} 1 & si\ s_i s_{i+g_2} \dots s_{i+g_p} = w_1 w_2 \dots w_p \\ 0 & sinon \end{cases} \tag{4.2}$$

Définition 6. *On dénote la* fréquence *d'un mot* $\mathbf{w}$ *dans une séquence* $\mathbf{S}$ *par*

$$N_{\mathbf{S}}(\mathbf{w}) = \sum_{i=0}^{n-k} Y_{\mathbf{S}}(i,\mathbf{w})$$

De manière analogue, la fréquence d'un mot espacé est donnée par

$$N_{\mathbf{S}}(\mathbf{w}^{\mathbf{g}}) = \sum_{i=0}^{n-l} Y_{\mathbf{S}}(i,\mathbf{w}^{\mathbf{g}})$$

Définition 7. *La* séquence de fréquences de mots *d'une séquence* $\mathbf{S}$ *est dénotée par*

$$F_{\mathbf{S}}(\mathbf{w}) = \{f_1,\dots,f_{n-k+1}\}\ où\ f_i = N_{\mathbf{S}}(\mathbf{w}_{\mathbf{S}}(i))\ pour\ tout\ i\ tel\ que\ 1 \le i \le n-k+1.$$

La séquence de fréquences de mots espacés *d'une séquence* **S** *est définie similairement par*

$$F_{\mathbf{S}}(\mathbf{w}^{\mathbf{g}}) = \{f_1, ..., f_{n-l+1}\} \text{ où } f_i = N_{\mathbf{S}}(\mathbf{w}^{\mathbf{g}}_{\mathbf{S}}(i)) \text{ pour tout } i \text{ tel que } 1 \leq i \leq n-l+1.$$

Définition 8. *La* table de fréquences *d'une séquence* **S**, *définie par*

$$Freq_{\mathbf{S}}(k) = \{(\mathbf{w}, N_{\mathbf{S}}(\mathbf{w}))\}, \forall \mathbf{w} \in \mathcal{A}^k$$

contient, pour chacun des mots de taille k possible, la fréquence totale de ce mot dans la séquence **S**.

Définition 9. *Soit* $\mathcal{N}_{\mathbf{S}}(k,f) = \{\mathbf{w} \in \mathcal{A}^k | N_{\mathbf{S}}(\mathbf{w}) = f\}$, *l'ensemble de mots de taille k d'une séquence* **S** *ayant une fréquence f. Le* spectrum *de taille k d'une séquence* **S**, *définit par*

$$Spec_{\mathbf{S}}(k) = \{|\mathcal{N}_{\mathbf{S}}(k,f)|\}, \forall f \in \mathbb{N}^+$$

contient, pour chacune des fréquences f, le nombre de mots ayant cette fréquence.

4.4 Survol de la solution proposée

Les définitions de la section précédente (Section 4.3) nous permettent de décrire la solution proposée. Pour annoter une séquence **S**, l'algorithme nécessite un ensemble de graines espacées $\mathcal{G} = \{g_1, g_2, \ldots, g_g\}$ de même poids p et une séquence d'entraînement annotée **T**. Puisqu'on possède le classificateur $H_T : \mathcal{I}_T \rightarrow \mathcal{Y}$, on peut obtenir la séquence d'entraînement non répétée $\mathbf{T}_{NR} = \{t_i | i \in \mathcal{I}_T \wedge H_T(i) = \text{Non Répétée}\}$ ainsi que la séquence d'entraînement répétée $\mathbf{T}_R = \{t_i | i \in \mathcal{I}_T \wedge H_T(i) = \text{Répétée}\}$.

La première étape consiste à estimer les paramètres initiaux du modèle de Markov caché avec les séquences d'entraînements $\mathbf{T}_{NR}$ et $\mathbf{T}_R$. Pour chacune de ces deux séquences, on calcule une table de fréquences par graine $g \in \mathcal{G}$. Un spectrum moyen par séquence est ensuite calculé à partir des tables de fréquence ; $Spec^*_{\mathbf{T}_{NR}}(p)$ et $Spec^*_{\mathbf{T}_R}(p)$. Ces deux spectrums sont utilisés pour calculer les paramètres des deux distributions double Pareto log-normale. Ces distributions seront ensuite mises à échelle en fonction de la longueur de la séquence à annoter et serviront de probabilités d'émission pour le modèle de Markov caché. On utilise des constantes comme

probabilités intitiales puisque celles-ci ont des effets marginaux sur les performances de la technique. Les probabilités de transition sont calculées à partir de compteur sur chacune des quatre transitions possibles ($NR \rightarrow NR$, $NR \rightarrow R$, $R \rightarrow NR$ et $R \rightarrow R$) rencontrées dans la séquence d'entraînement. Le modèle de Markov est schématisé à la Figure 4.2.

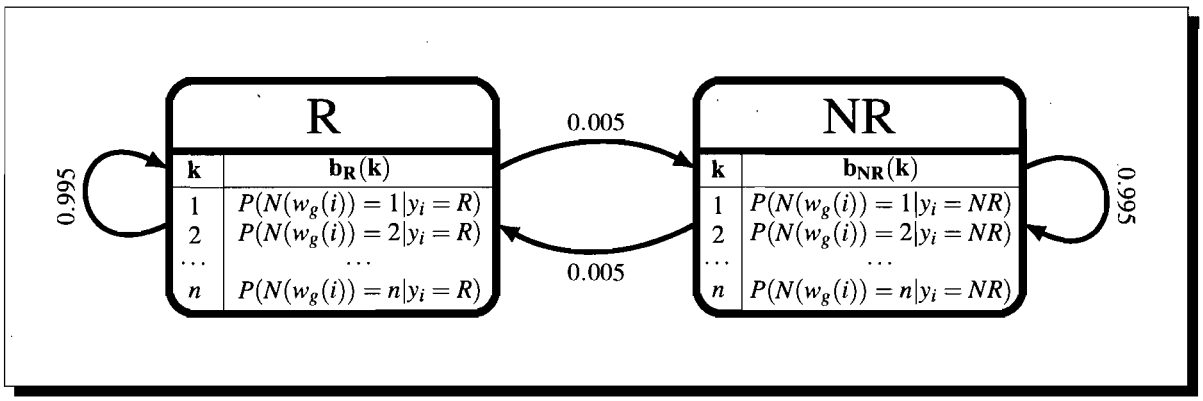


Figure 4.2: Représentation graphique du modèle de Markov caché utilisé pour l'annotation de régions répétées.

Une fois les paramètres initiaux du modèle de Markov caché trouvés, on peut utiliser le modèle pour annoter une séquence cible **S**. On calcule d'abord la séquence de fréquence de **S** pour chaque graine $g \in \mathcal{G}$. On calcule ensuite, pour chaque position d'une séquence de fréquences donnée, la probabilité d'étiquetage a postériori, c.-à-d. la probabilité d'une étiquette étant donné la séquence et le modèle. Une fois ce processus effectué sur chacune des séquences de fréquences, on normalise les probabilités à postériori par le nombre de graines espacées utilisées. On associe ensuite à chaque position l'étiquette ayant la probabilité moyenne la plus élevée.

CHAPITRE 5

ALGORITHME

Les modèles de Markov cachés sont des modèles probabilistes dans lesquels on présuppose que le système modélisé est un processus Markovien ; c.-à-d. la distribution conditionnelle des probabilités des états futurs ne dépend que de l'état présent et d'un certain nombre d'états passés. Le modèle est dit "caché" puisque la séquence d'états n'est pas directement observable. On observe plutôt la séquence de caractères émise par les états. Puisque chaque état possède sa propre distribution de probabilité d'émission de caractères, il est possible d'inférer la séquence d'états à partir de la séquence de caractères observés. Notre application utilise un modèle de Markov caché dans lequel les états représentent les régions répétées et non répétées. Les distributions de probabilité d'émission des deux états sont données par la distribution double Pareto log-normale calculée à partir des spectrums des régions répétées et non répétées, respectivement. La séquence observée est la séquence de fréquence telle que décrite dans la définition 7 de la section 4.3. Nous désirons donc, à partir de la séquence de fréquences de la séquence cible **S**, découvrir la séquence d'états sous-jacente la plus probable, c.-à-d. l'annotation des régions répétées.

Définition 10. *Pour une séquence cible donné S et un un mot espacé* $\mathbf{w^g}$ *on définit la séquence d'observation du modèle de Markov caché par*

$$F_{\mathbf{S}}(\mathbf{w^g}) = \{f_1, ..., f_{n-l+1}\}$$

tel que décrit dans la définition 7 de la section 4.3. On peut maintenant définir de manière formelle le modèle de Markov caché utilisé dans l'application par le quintuplet $\langle \mathcal{S}, \mathcal{V}, \Pi, A, B \rangle$ *où :*

- $\mathcal{S}$ *représente l'ensemble d'états :*

$$\mathcal{S} = \{S_1, S_2\} = \{R, NR\}$$

où R=Répétée, NR=Non Répétée. De plus, on définit par

$$Q = q_1 \dots q_n, \quad q_i \in \mathcal{S}$$

la séquence d'états du modèle de Markov caché.

◇ $\mathcal{V}$ *représente l'alphabet :*

$$\mathcal{V} = \{v_1, \dots, v_m\}$$

où $v_i \in \mathcal{V}$ *est une fréquence de mot.*

◇ $\Pi = \mathcal{S} \rightarrow [0,1]$ *représente les probabilités d'état initial :*

$$\Pi = \{\pi_i\} = \{P(q_1 = i)\}, \quad 1 \leq i \leq 2.$$

◇ $A = \mathcal{S} \times \mathcal{S} \rightarrow [0,1]$ *représente la matrice des probabilités de transition ; on dénote par* a_{ij} *la probabilité de transition de l'état i à l'état j :*

$$A = \{a_{ij}\} = \{P(q_t = S_j | q_{t-1} = S_i)\}, \quad 1 \leq i, j \leq 2.$$

◇ $B = \mathcal{S} \times \mathcal{V} \rightarrow [0,1]$ *représente la matrice des probabilités d'émission ; on dénote par* $b_i(k)$ *la probabilité d'émettre le caractère* v_k *à l'état i :*

$$B = \{b_i(k)\} = \{P(f_t = v_k | q_t = S_i)\}, \quad 1 \leq i \leq 2, \ 1 \leq k \leq m.$$

5.1 Estimation des paramètres initiaux

La première étape consiste à estimer les paramètres du modèle de Markov caché, c.-à-d. les probabilités d'état initial (Π), les probabilités de transition (A) et les probabilités d'émission (B), et ce, pour chacun des deux états. Cette étape nécessite une séquence d'ADN d'entraînement T dans laquelle les régions répétées sont préalablement annotées ainsi qu'un ensemble $\mathcal{G} = \{g_1, \dots, g_m\}$ de m graines espacées de même poids p. On définit par T_{NR} et T_R la séquence d'entraînement dans laquelle les régions répétées et non répétées ont été masquées, respectivement.

5.1.1 Estimation des probabilités d'état initial (Π)

Les probabilités d'état initial sont estimées par la distribution de probabilités stationnaires (π_{NR}, π_R) statisfaisant les conditions

$$\sum_{i \in \{R,NR\}} \pi_i = 1$$

et

$$\pi_j = \sum_{i \in \{R,NR\}} \pi_i a_{ij} \quad \text{pour} j \in \{R,NR\}$$

En résolvant ces équations linéaires, on obtient

$$\pi_{NR} = \frac{a_{R \to NR}}{a_{NR \to R} + a_{R \to NR}} \quad \text{et} \quad \pi_R = \frac{a_{NR \to R}}{a_{NR \to R} + a_{R \to NR}}$$

5.1.2 Calcul des probabilités de transitions (A)

Rappelons que l'on définit l'ensemble d'indices de la séquence d'entraînement $\mathbf{T} = t_1 \ldots t_m$ par $\mathfrak{I}_T = \{1, \ldots, m\}$. Soit $Trans_i(a,b)$ la fonction indicatrice de transition définie par :

$$Trans_i(a,b) = \begin{cases} 1 & \text{si } H_T(i) = a \text{ et } H_T(i+1) = b; \quad a,b \in \mathcal{Y}, \quad i, i+1 \in \mathfrak{I}_T \\ 0 & \text{sinon} \end{cases} \tag{5.1}$$

alors, on peut formellement définir la probabilité de transition de la région a à la région b par

$$a_{a \to b} = \frac{\sum_{i=0}^{n-1} Trans_i(a,b)}{n-1}$$

Puisque $\mathbf{T}$ est annoté, le calcul des probabilités de transition est élémentaire. La séquence est parcourue, et, pour toute paire de nucléotides consécutives (t_i, t_{i+1}), le compteur correspondant à la transition utilisée est incrémenté. Finalement, les compteurs sont divisés par le nombre total de transitions pour donner la probabilité de transition.

5.1.3 Calcul des probabilités d'émission (B)

Le modèle de Markov caché nécessite des probabilités d'émission pour chacun des deux états. Le calcul des probabilités d'émission s'effectue en quatre phases ; le calcul des fré-

quences d'oligonucléotides (5.1.3.1), le calcul du spectrum (5.1.3.2), le calcul de la distribution double Pareto log-normale (5.1.3.3) et, finalement, l'ajustement des distributions DPLN (5.1.3.4).

5.1.3.1 Calcul de la table de fréquences d'oligonucléotides

La table de fréquences d'oligonucléotides, telle que décrite dans la définition 8 de la section 4.3, possède 4^p clés correspondant aux 4^p oligonucléotides de taille p. À chaque clé de la table est associé le nombre d'occurrences de l'oligonucléotide correspondant trouvées dans la séquence. On calcule une table de fréquences d'oligonucléotides pour chacune des deux régions R et NR et pour chacune des graines espacées $g \in \mathcal{G}$, de manière à éviter le débordement d'entier (*integer overflow*). Pour une graine et une région données, on parcourt la séquence masquée et on incrémente la clé de la table correspondant à l'oligonucléotide rencontré à chaque position. L'algorithme 1 illustre cette procédure.

Algorithme 1 : Calcul de la table des fréquences

Entrées : Séquence d'entraînement T,Annotation A, Ensemble de graines espacées $\mathcal{G}$

Sorties : La table des fréquences $frequences$

```
pour chaque e ∈ {R,NR} faire
    T_e = masqueSequence(T,A,e) ;
    pour chaque g ∈ G faire
        pour pos = 1 à |T_e| − l + 1 faire
            frequences[e][g][w^g_{T_e}(pos)] += 1 ;
        fin
    fin
fin
retourner frequences
```

5.1.3.2 Calcul du spectrum génomique

Le spectrum (Definition 9, Section 4.3) contient comme clé une fréquence d'oligonucléotides et comme valeur, le nombre d'oligonucléotides de taille p ayant cette fréquence. Pour

calculer cette table, on parcourt chacune des tables de fréquences d'oligonucléotides, et, pour chaque valeur rencontrée, on incrémente la clé correspondante dans le spectrum. L'algorithme 2 présente cette procédure.

Algorithme 2 : calcul du spectrum

Entrées : Table des fréquences $frequence$

Sorties : Le spectrum $spectrum$

```
pour chaque e ∈ {R,NR} faire
    pour chaque g ∈ 𝒢 faire
        pour chaque w ∈ 𝒜^p faire
            spectrum[e][frequences[e][g][w]] += 1 ;
        fin
    fin
fin
retourner spectrum
```

Un exemple de table de fréquences d'oligonucléotide et de table de spectre génomique est donné à la Figure 5.1.

5.1.3.3 Calcul de la distribution Double Pareto Log-Normale

Les paramètres de la distribution double Pareto log-normale sont inférés par ajustement de courbe (*curve fitting*) en utilisant la méthode décrite dans [66].

5.1.3.4 Ajustement de la distribution DPLN

Le paramètre ν, qui détermine le "centre" de la distribution DPLN, dépend fortement du nombre de mots contenus dans la séquence observée (Figure 3.11, bas). Puisque les nombres de mots contenus dans les séquences d'entraînement et dans la séquence cible peuvent différer, il est primordial d'ajuster les probabilités d'émission pour tenir compte de cet écart. On veut donc ajuster le paramètre ν de la distribution DPLN au nombre de mots contenus dans la séquence cible.

(a) Seq : CCATCAGCACTCCCAGTTCCTTTTCCTTTCTCTGCATGAAGTGTTTCATC
Graine espacée : 1001

$\mathbf{w}^g$	$N_S(\mathbf{w}^g)$
AA	3
AC	1
AG	1
AT	2
CA	1
CC	5
CG	2
CT	8
GA	0
GC	2
GG	1
GT	3
TA	3
TC	7
TG	2
TT	6

(b)

$N_S(\mathbf{w}^g)$	$\lvert\mathcal{N}_S(p, N_S(\mathbf{w}^g))\rvert$
0	1
1	4
2	4
3	3
4	0
5	1
6	1
7	1
8	1

(c)

Figure 5.1: Table de fréquences d'oligonucléotides (b) et table de spectre génomique (c) produite en parcourant la séquence CCATCAGCACTCCCAGTTCCTTTTCCTTTCTCTGCATGAAGTGTTTCATC avec la graine espacée 1001 (a).

On définit les facteurs d'ajustements par

$$adj_e = \frac{|S|}{|T_e|} \quad \text{pour } e \in \{R, NR\}$$

où $|S|$ et $|T_e|$ représentent le nombre de mots contenus dans la séquence cible S et dans la séquence d'entrainement T_e, respectivement. L'espérance du nombre d'occurrences de mots u est définie par la formule [66]

$$E(u) = \frac{\alpha\beta}{(\alpha-1)(\beta+1)} e^{\nu+\frac{\tau^2}{2}}$$

Pour une région donnée $e \in \{R, NR\}$, on estime donc le nouveau coefficient $\hat{\nu}$ par :

$$\frac{\alpha\beta}{(\alpha-1)(\beta+1)} e^{\hat{\nu}+\frac{\tau^2}{2}} = adj_e \frac{\alpha\beta}{(\alpha-1)(\beta+1)} e^{\nu+\frac{\tau^2}{2}}$$

$$e^{\hat{\nu}} = adj_e \times e^{\nu}$$

$$\hat{\nu} = ln(adj_e) + \nu$$

La Figure 5.2 affiche un exemple d'ajustement de distributions de probabilités d'émission.

5.2 Décodage a postériori

Le décodage a postériori consiste à associer une étiquette à chaque position de la séquence cible en fonction de la probabilité a postériori moyenne. La probabilité a postériori d'une étiquette à une position donnée est la somme des probabilités normalisées de tous les étiquetages possibles ayant cette étiquette à cette position, c.-à-d la probabilité d'avoir cette étiquette à une position donnée dans la séquence, étant donné la séquence et le modèle. On doit donc, pour chaque graine espacée, calculer les probabilités a postériori d'être dans une région répétée pour chaque position de la séquence. Pour une position donnée, on calcule la probabilité a postériori moyenne ; si cette probabilité moyenne est supérieure à un seuil donné, on étiquette cette position comme faisant partie d'une région répétée. L'algorithme 3 résume cette procédure.

Aux lignes 1 à 6, on calcule les probabilités a postériori moyennes d'être dans une région répétée pour chaque position de la séquence. À la ligne 2, on fait appelle à la fonction définit dans l'algorithme 4 pour calculer les probabilité a postériori de la région répétée (R) pour une

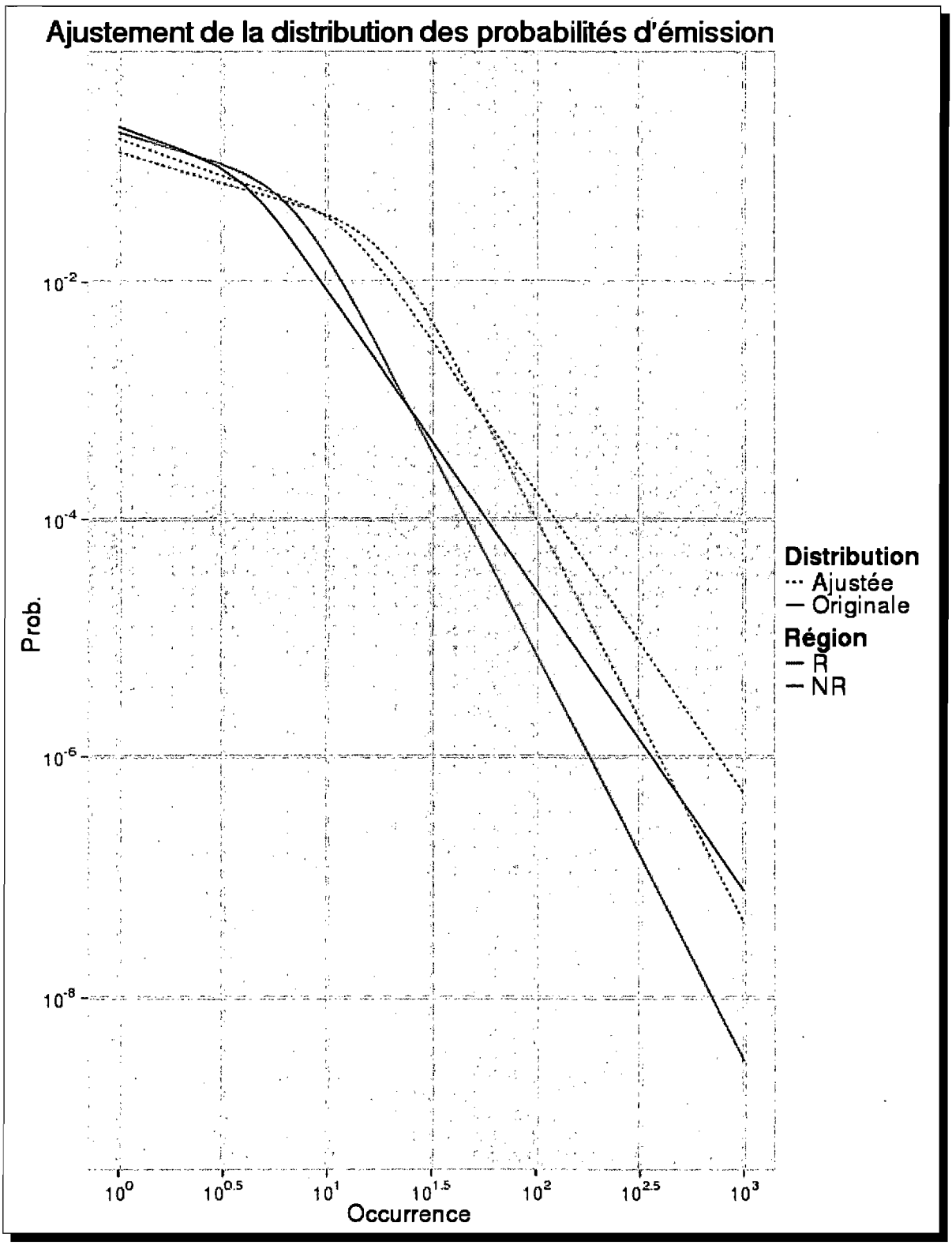


Figure 5.2: Exemple d'ajustement des distributions de probabilités d'émission. Le seul paramètre modifié est ν, ce qui explique que les queues soient parallèles.

graine donnée. Aux lignes 7 à 13, on étiquette chaque position de la séquence en fonction des probabilités a postériori moyennes. La sous-section 5.2.1 explique le calcul des probabilités a postériori.

Algorithme 3 : Décodage à posteriori

Entrées : Séquence d'ADN S, Ensemble de graines espacées *graines*, Hmm *hmm*, Valeur réelle *seuil*

Sorties : L'étiquettage *etiquette*

```
   /* Calcule des probabilités a postériori moyennes          */
1  pour chaque g ∈ 𝒢 faire
2  |  posterior ← probabiliteAPosteriori(hmm, S, g);
3  |  pour pos = 1 à |S| faire
4  |  |  prob[pos] += posterior[pos] / |𝒢|;
5  |  fin
6  fin
   /* Étiquettage des positions                               */
7  pour pos = 1 à |S| faire
8  |  si prob[pos] > seuil alors
9  |  |  etiquette[pos] = R
10 |  sinon
11 |  |  etiquette[pos] = NR
12 |  fin
13 fin
14 retourner etiquette
```

5.2.1 Calcul des probabilités a postériori

Le calcul des probabilités a postériori de la séquence d'observations O étant donné le modèle $\lambda = \langle \Pi, A, B \rangle$ nécessite deux procédures appelées *forward* et *backward* présentées dans les sous-sections 5.2.1.1 et 5.2.1.2 respectivement. La sous-section 5.2.1.3 présente la procédure utilisée pour calculer les probabilités a postériori.

5.2.1.1 Procédure forward

Considérons la variable *forward* $\alpha_t(i)$ définie par :

$$\alpha_t(i) = P(f_1 f_2 \dots f_t, q_t = S_i|\lambda)$$

c.-à-d., la probabilité de générer la séquence d'observation partielle $f_1 f_2 \dots f_t$ et d'être arrivé à l'état S_i à l'instant t, étant donné le modèle $\lambda = \langle \Pi, A, B \rangle$. On peut calculer $\alpha_t(i)$ avec la récursion suivante :

1. Initialisation :

$$\alpha_1(i) = \pi_i b_i(f_1), 1 \leq i \leq 2.$$

2. Induction :

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{2} \alpha_t(i) a_{ij}\right] b_j(f_{t+1}), \quad 1 \leq j \leq 2, \quad 1 \leq t \leq T-1.$$

3. Terminaison :

$$P(O|\lambda) = \sum_{i=1}^{2} \alpha_T(i)$$

On appelle cet algorithme "*forward*" puisque l'induction est réalisée vers l'avant : on calcule d'abord la probabilité de générer le premier symbole, puis on ajoute un symbole à la fois jusqu'à ce que tous les symboles aient été ajoutés. Un algorithme similaire présenté dans la prochaine section est utilisé pour réaliser ce calcul à l'envers.

5.2.1.2 Procédure backward

De manière analogue, on définit la variable *backward* $\beta_t(i)$ par :

$$\beta_t(i) = P(f_{t+1} f_{t+2} \dots f_T | q_t = S_i, \lambda)$$

qui représente la probabilité de générer la sous-séquence d'observation $f_{t+1} f_{t+2} \dots f_T$ en partant de l'état S_i au temps t et en utilisant le model λ. On peut encore une fois calculer $\beta_t(i)$ à l'aide de récursion :

1. Initialisation :

$$\beta_T(i) = 1, \quad 1 \leq i \leq 2.$$

2. Induction :

$$\beta_t(j) = \sum_{i=1}^{2} a_{ij} b_j(f_{t+1}) \beta_{t+1}(j), \quad 1 \leq i \leq 2, \quad t = T-1, T-2, \ldots, 1.$$

3. Terminaison :

$$P(O|\lambda) = \sum_{i=1}^{2} \alpha_T(i)$$

5.2.1.3 Probabilité a postériori

Les variables forward-backward permettent de calculer la probabilité de se trouver à l'état S_i à l'instant t, en générant la séquence d'observation O avec les paramètres λ. Cette probabilité est définie par la variable $\gamma_t(i)$, c.-à-d.

$$\gamma_t(i) = P(q_t = S_i|O, \lambda)$$

Puisque la variable $\alpha_t(i)$ exprime la probabilité de générer la séquence $f_1 f_2 \ldots f_t$ et d'être à l'état S_i au temps t, et puisque la variable $\beta_t(i)$ exprime la probabilité de générer la séquence $f_{t+1} f_{t+2} \ldots f_T$ en débutant à l'état S_i au temps t, la variable $\gamma_t(i)$ peut être exprimée par

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{2} \alpha_t(j)\beta_t(j)}$$

Le facteur de normalisation $P(O|\lambda) = \sum_{j=1}^{2} \alpha_t(j)\beta_t(j)$ permet d'avoir

$$\sum_{i=1}^{2} \gamma_t(i) = 1$$

L'algorithme 4 calcule les probabilités a postériori d'être dans une région répétée pour une graine espacée donnée et une séquence donnée. Aux lignes 1 à 3, on calcule la table de fréquences (Définition 8, Section 4.3). Aux lignes 4 à 11, on calcule les valeurs αs tel que définis dans la section 5.2.1.1 alors que les valeurs βs définis dans la section 5.2.1.2 sont calculées aux lignes 12 à 19. On utilise la séquence de fréquences de mots espacés (Définition 7, Section 4.3) de la séquence cible comme valeurs observées. Finalement, les probabilités à postériori pour la région répétée sont calculées aux ligne 20 à 22.

Algorithme 4 : ProbabiliteAPosteriori

Entrées : Hmm *hmm*, Séquence d'ADN S, Graine espacée g

Sorties : Probabilité a postériori de l'état R *posterior*

/* Calcul des fréquences */

pour $t = 1$ à $|S| - l + 1$ faire

  $frequences[w_S^g(t)] += 1$;

fin

/* Calcul des valeurs αs */

pour chaque $e \in \{R, NR\}$ faire

  $\alpha_1(e) = \pi_e \times b_e(frequences[w_S^g(1)])$

fin

pour $t = 1$ à $|S| - l$ faire

  pour chaque $e_2 \in \{R, NR\}$ faire

    $$\alpha_{t+1}(e_2) = \left[\sum_{e_1 \in \{R,NR\}} \alpha_t(e_1) \times a_{e_1 e_2} \right] \times b_{e_2}(frequences[w_S^g(t+1)])$$

  fin

fin

/* Calcul des valeurs βs */

pour chaque $e \in \{R, NR\}$ faire

  $\beta_{|S|-l+1}(e) = 1$

fin

pour $t = |S| - l$ à 1 faire

  pour chaque $e_2 \in \{R, NR\}$ faire

    $$\beta_t(e_2) = \sum_{e_1 \in \{R,NR\}} a_{e_1 e_2} \times b_{e_2}(frequences[w_S^g(t+1)]) \times \beta_{t+1}(e_2)$$

  fin

fin

/* Calcul des valeurs a postériori */

pour $t = 1$ à $|S| - l + 1$ faire

  $$posterior[t] = \frac{\alpha_t(R) \times \beta_t(R)}{\sum_{e \in \{R,NR\}} \alpha_t(e) \times \beta_t(e)}$$

fin

retourner *posterior*

5.3 Complexité algorithmique

Soit g le nombre de graines espacées de poids p, et soit n et m le nombre de mots contenus dans la séquence cible et dans la séquence d'entraînement, respectivement. Le calcul de la table de fréquences d'oligonucléotides nécessite le parcours de la séquence d'entraînement pour chaque graine espacée et pour chaque état. On a donc $2 * g * m$ opérations $\sim O(gm)$. Le calcul du spectrum requiert une opération pour chaque mot de taille p distinct, pour chaque graine et chaque état. La complexité est donc de $O(2g4^p) \sim O(g4^p)$. L'entraînement du modèle de Markov caché possède donc une complexité de $O(g * \text{Max}(4^p, m))$. Pour les chromosomes du génome humain, $m > 4^p$ si $p \leq 12$.

Le calcul des probabilités a postériori pour une graine espacée donnée nécessite $O(n)$ opérations. Le calcul des probabilités a postériori moyenne nécessite aussi $O(n)$ opérations. Puisque ces opérations sont effectuées pour chacune des g graines espacées, on obtient donc $O(gn)$ opérations. Finalement, n opérations sont nécessaires pour l'étiquettage de la séquence. La complexité de la phase de décodage est donc de $O(gn)$. La complexité totale de l'algorithme est donc linéaire en fonction de la taille des séquences.

CHAPITRE 6

RÉSULTATS

Nous avons comparé les annotations produites par notre algorithme à celles produites par RepeatMasker [73]. Cette application génère, de l'avis de tous, les meilleures annotations de régions répétées en utilisant une méthode très lente basée sur l'alignement de séquences.

6.1 Mesures de performance

Pour l'analyse de résultats binaires, deux mesures de performance statistique sont généralement utilisées : la sensibilité et la spécificité. La sensibilité mesure la capacité de l'algorithme à annoter les régions comme répétées lorsqu'elles le sont réellement. Par opposition, la spécificité est l'aptitude de l'algorithme à ne pas annoter les régions comme répétées lorsqu'elles ne le sont pas. En se référant au tableau , on définit formellement la sensibilité par :

$$\text{Sensibilité} = \frac{\text{Vrai Positif}}{\text{Vrai Positif} + \text{Faux Négatif}}$$

alors que la spécificité est définie par :

$$\text{Spécificité} = \frac{\text{Vrai Négatif}}{\text{Vrai Négatif} + \text{Faux Positif}}$$

On aimerait que ces deux mesures soient le plus élevées possible. Or, en pratique, elles sont souvent inversement proportionelles ; plus la sensibilité est élevée, plus la spécificité est faible et vice-versa. De fait, il est donc parfois hasardeux d'utiliser ces deux seules mesures pour estimer les performances d'un algorithme. Nous utiliserons donc une troisième mesure de per-

		Réel	
		Répétées	Non Répétées
Estimé	Réptées	Vrai Positif	Faux Positif
	Non Répétées	Faux Négatif	Vrai Négatif

Tableau 6.I: Résultats possibles lors de l'annotation. Les rangées représentent les nucléotides annotés par l'algorithme alors que les colonnes représentent les nucléotides annotés par *RepeatMasker* [73]

formance tirée du domaine de la recherche d'information appelée mesure-F (*F-mesure*). Dans ce domaine, les critères de mesure de performances utilisés sont la sensibilité et la précision. La précision est le nombre de nucléotides correctement annotés comme répétés divisé par le nombre total de nucléotides annotés comme répétés par l'algorithme :

$$\text{Précision} = \frac{\text{Vrai Positif}}{\text{Vrai Positif} + \text{Faux Positif}}$$

On définit la mesure-F par :

$$F_\beta = \frac{(1+\beta^2) \times \text{Précision} \times \text{Sensibilité}}{\beta^2 \times \text{Précision} + \text{Sensibilité}}$$

Dans cette étude, nous utilisons $\beta = 0.5$ puisque nous désirons accorder plus d'importance à la précision qu'à la sensibilité. Nous estimons qu'il est moins négligeable d'omettre une région répétée que d'en annoter une fictive, puisque dans le premier cas, l'annotation est incomplète alors que dans le second, elle est erronée. La mesure-F est donc définie par :

$$F_{0.5} = \frac{1.25 \times \text{Vrai Positif}}{(1.25 \times \text{Vrai Positif}) + (0.25 \times \text{Faux Négatif}) + \text{Faux Positif}}$$

6.2 Influence du nombre de graines espacées

Dans le premier test, nous avons effectué l'annotation du chromosome humain 12 en entraînant notre modèle de Markov caché avec le chromosome 11. Nous avons utilisé jusqu'à 10 graines espacées de poids 12 (Figure 6.1). L'utilisation d'un plus grand nombre de graines espacées permet d'améliorer la spécificité au détriment de la sensibilité. Cependant, puisque l'augmentation globale de spécificité est supérieure à la diminution globale de sensibilité, on peut conclure qu'une augmentation du nombre de graines augmente légèrement les performances de la recherche. Il est à noter que la différence de performance n'est pas linéaire ; elle est beaucoup plus marquée lorsque le nombre de graines est petit ($g \leq 3$). Ces résultats sont conformes à la théorie des graines espacées. Malgré le fait que les graines de même poids possèdent la même espérance en terme de nombre de *hit*, elle n'ont pas la même sensibilité [50]. Cette différence de sensibilité est attribuable au nombre de chevauchements ; moins il y a de chevauchements entre une graine à une position p et cette même graine à une position $p+1$, meilleure est la sensibilité. Un bon ensemble de graines doit minimiser le nombre de

chevauchements entre toutes les graines le composant. Or, plus on ajoute de graines, plus les chevauchements sont fréquents et plus le gain de sensibilité est faible. À la lumière de ces résultats, il semble que l'utilisation de 3 graines espacées soit suffisante.

6.3 Influence du poids des graines espacées

Dans le deuxième test, nous avons encore une fois annoté le chromosome humain 12 en l'entraînant avec le chromosome 11 (Figure 6.2). L'annotation effectuée avec des graines de poids inférieurs à 11 s'avère inéfficace. Cela s'explique en grande partie par le fait que, pour ces poids de graines, la queue gauche de la distribution double Pareto log-normale s'ajuste mal aux spectrums (voir Figure 3.9, section 3.4.3.1). Par contre, pour des poids supérieurs à 10, la recherche est relativement efficace. En effet, pour cet intervalle de poids, la sensibilité augmente d'environ 24% alors que la spécificité diminue d'environ 10%. On constate aussi que l'influence du poids de la graine est considérablement supérieure à celle du nombre de graines.

6.4 Influence de l'ajustement de la distribution DPLN

La séquence d'entraînement et la séquence cible utilisées dans les tests des deux sections précédentes ont une taille relativement similaire. Dans ce test nous annotons le chromosome 12 du génome humain en entraînent l'algorithme avec des chromosomes humains de tailles différentes. La Figure 6.3 affiche les performances d'annotation. Des résultats moyens de 65.51%, 82.24% et 47.88% avec écart type de 3.02%, 8.1% et 8.69% on été obtenus pour la mesure-F, la spécificité et la sensibilité respectivement. Si on omet le chromosomes 21 pour lequel les résultats sont plus faibles, les écarts types de la spécificité et de la sensibilité tombent a 4.44% et 5.33% respectivement. Si on tient compte du fait que les différences de performance sont aussi dues aux variations en éléments répétés des séquences d'entraînements, il semble que l'ajustement de la distribution DPLN soit efficace.

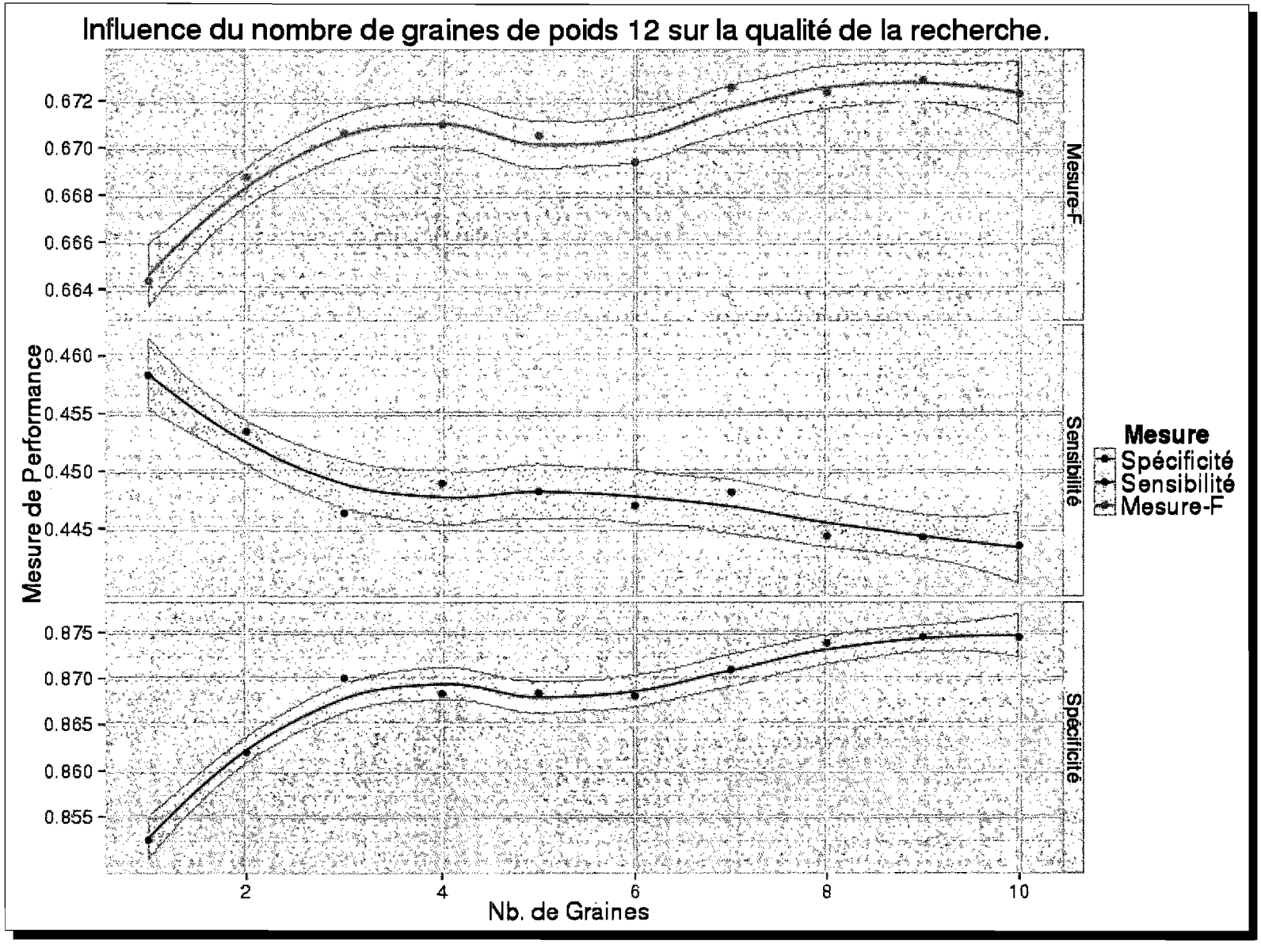


Figure 6.1: Influence du nombre de graines espacées de poids 12 sur la qualité de la recherche.

6.5 Influence de l'organisme

Dans ce test, nous avons voulu vérifier la qualité de l'annotation du chromosome humain 12 effectuée par l'algorithme en l'entraînant avec des séquences d'entraînement provenant de différents organismes relativement proches de l'homme. La Figure 6.4 affiche un histogramme des différentes mesures en fonction de la séquence d'entraînement utilisée. À notre grande surprise, les performances sont relativement élevées avec des résultats moyens de 66.65%, 89.84% et 40.80% pour la mesure-F, la spécificité et la sensibilité respectivement. De manière générale, les résultats sont plus conservateurs (spécificité plus élevée, sensibilité plus faible) que ceux obtenus dans les autres tests. De tous les tests effectués dans ce chapitre, la mesure-F et la spécificité maximale (71.66% et 95.85%, respectivement) ont été atteintes en entraînant le chromosome 12 avec le chromosome 1 du génome du poisson zèbre (*zebrafish*). Ces résultats sont encourageants puisqu'ils impliquent qu'il est possible d'utiliser notre technique sur des organismes nouvellement séquencés n'ayant aucune annotation de régions répétées en utilisant des séquences d'entraînement d'organismes relativement proches.

6.6 Entraînement Baum-Welch

Nous avons tenté d'entraîner les paramètres de l'HMM avec une version modifiée de l'algorithme Baum-Welch [3]. Les probabilités d'état initial et les probabilités de transition sont recalculées selon la technique standard. Les probabilités d'émission sont réestimées en utilisant les probabilités à postériori. Pour ce faire, on recalcule les tables de fréquences de la manière suivante : pour une région $e \in R, NR$ et pour chaque position i de la séquence, on incrémente l'entrée de la table de fréquences correspondant au mot espacé débutant à la position i par la probabilité à postériori d'être dans l'état e à cette position i. On utilise ensuite ces tables pour calculer le spectrum (section 5.1.3.2) et la distribution DPLN (section). Dans les faits, cette procédure peut-être réitérée un certain nombre de fois jusqu'à ce que l'algorithme atteigne un optimum local. En pratique, pour une raison que nous ignorons, cette technique d'entraînement s'est avérée très instable. D'une itération à l'autre, on obtient parfois une augmentation, parfois une diminition des performances d'annotation. Puisqu'il est impossible à priori de savoir quand l'entraînement doit être arrêté pour maximiser les performances, nous avons décidé de ne pas utiliser cette technique, d'autant plus qu'elle est coûteuse en temps.

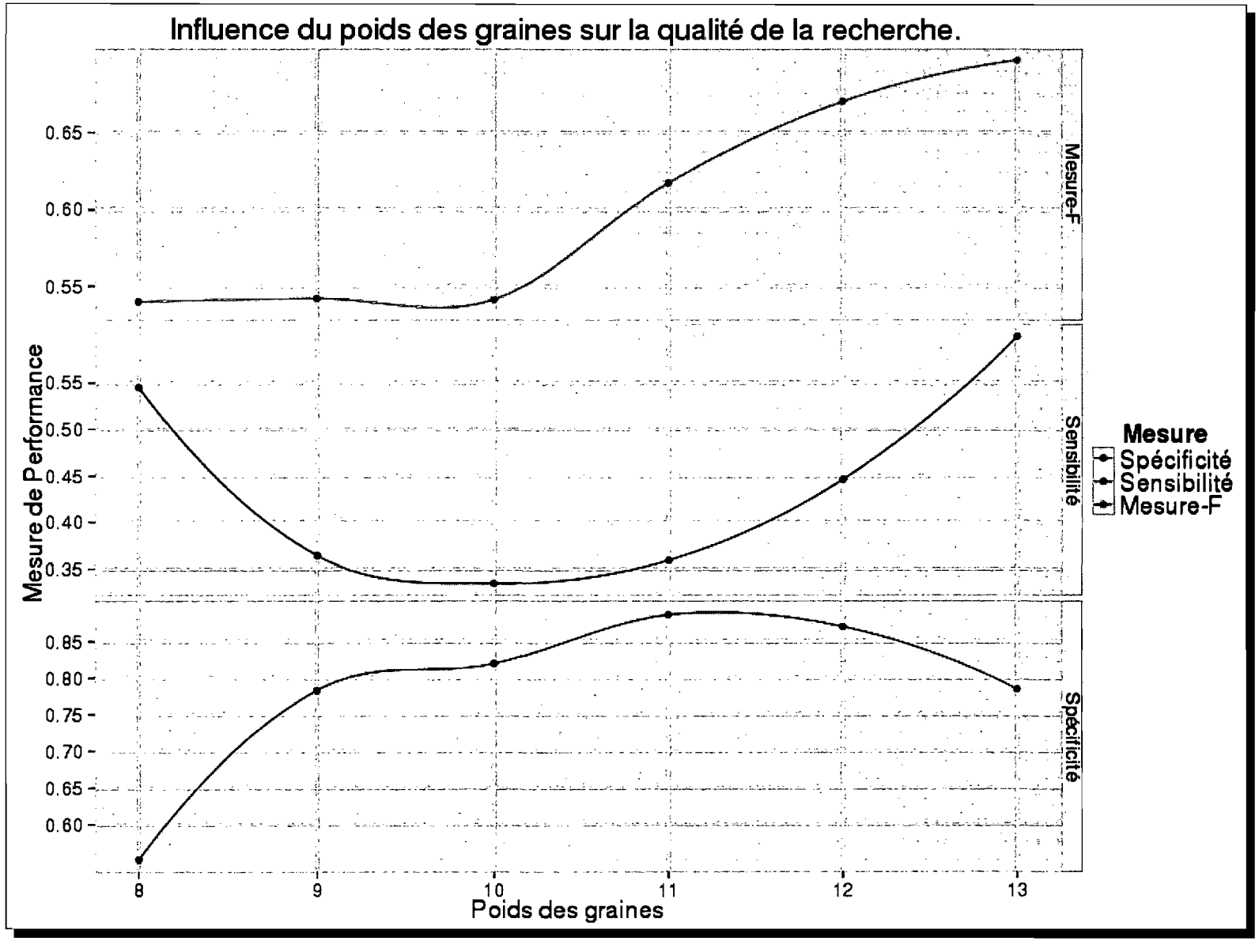


Figure 6.2: Influence du poids des graines espacées sur la qualité de la recherche, en utilisant 3 graines espacées.

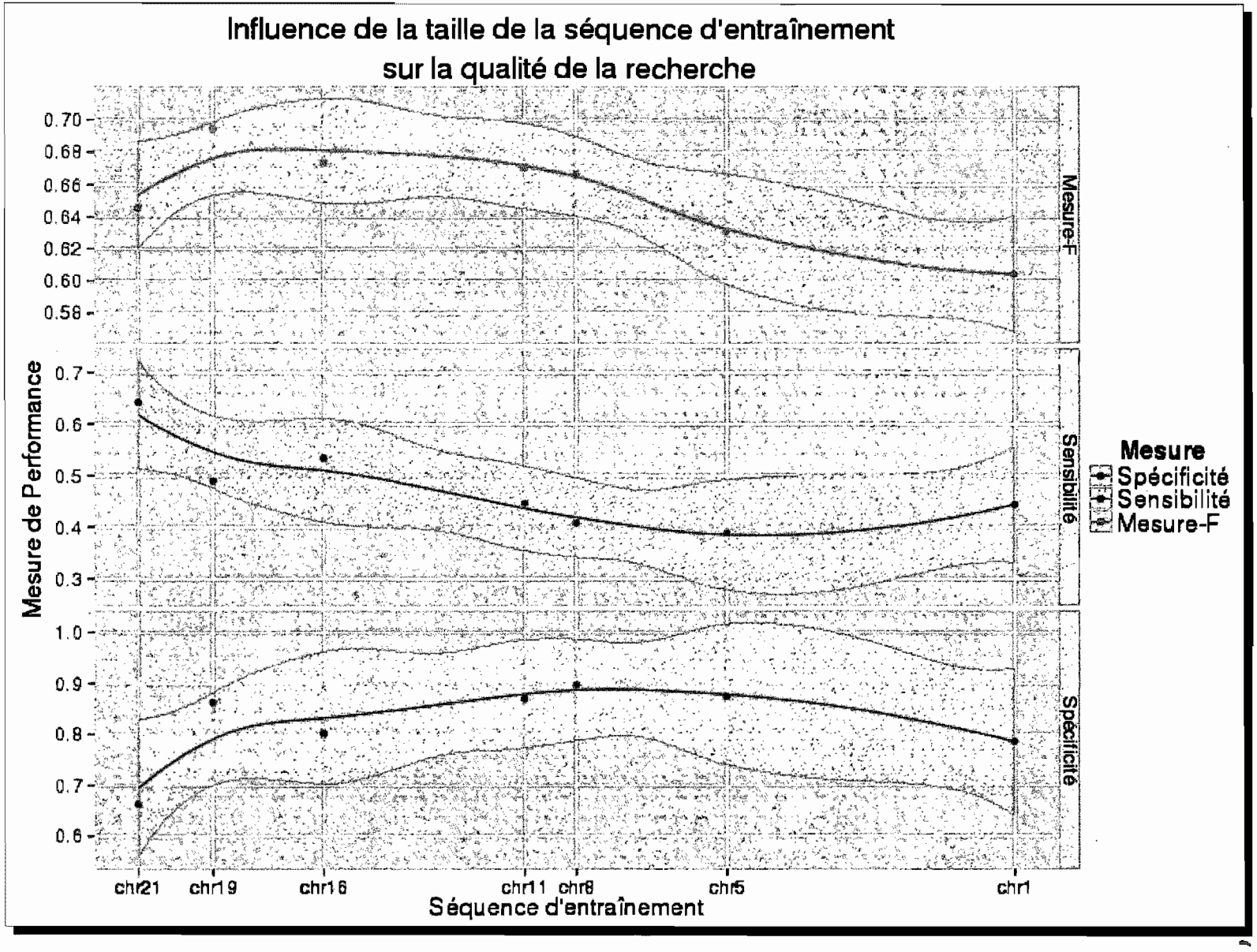


Figure 6.3: Influence de la taille de la séquence d'entraînement sur la qualité de la recherche. Afin de valider la technique d'ajustement de la distribution DPLN, le chromosome 12 a été entraîné avec des séquences d'entraînement de différentes tailles en utilisant 3 graines espacées de poids 12.

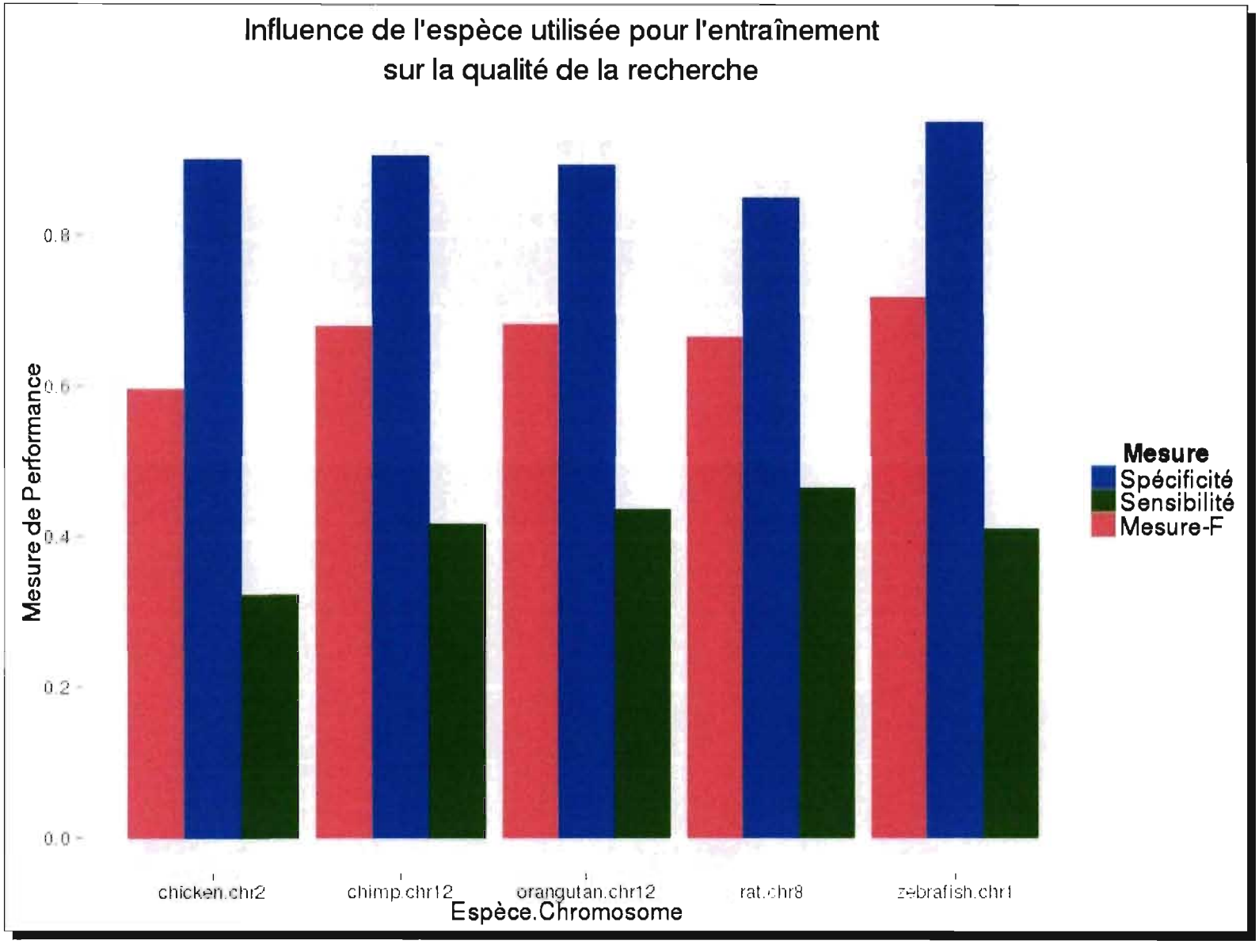


Figure 6.4: Influence de l'origine de la séquence d'entraînement sur la qualité de la recherche. Le chromosome humain 12 a été entraîné avec des séquences d'entraînement provenant de différentes espèces en utilisant 3 graines espacées de poids 12.

CONCLUSION

6.7 Conclusion

Nous avons développé une technique d'annotation de régions répétées De Novo élégante utilisant une distribution d'occurrences de mots suivant un loi double Pareto log-normale comme modèle nul. Malgré des résultats modestes par rapport aux annotations générées par RepeatMasker, nous avons démontré que l'idée n'est pas aberrante.

Nous avons d'abord prouvé qu'il est possible d'atteindre une sensibilité de 45% tout en conservant une spécificité supérieur à 80% en utilisant 3 graines espacées de poids 12 ou 13. Nous avons de plus illustré qu'il est possible, en ajustant la distribution double Pareto log-normale, d'entraîner la chaîne de Markov avec des séquences de tailles distinctes sans trop diminuer les performances. Finalement nous avons établi qu'il est possible d'entraîner le modèle de Markov caché avec des séquences d'entraînements provenant d'organismes différents.

6.8 Perspectives

L'amélioration la plus cruciale consiste à développer une technique d'entraînement stable. Les chaînes de Markov caché bénéficient généralement d'un bon gain de performance lorsqu'elles sont entraînées de manière adéquate.

Une seconde piste à explorer serait l'utilisation d'une chaîne de Markov d'ordre supérieur. On pourrait potentiellement extraire plus d'informations des séquences de fréquences et ainsi améliorer les annotations produites par l'algorithme. Cependant, le nombre d'émission relativement élevé limite cette approche puisque cela impliquerait une plus grande utilisation de mémoire.

Finalement, puisque comme on l'a indiqué dans la section 6.5, il est possible d'entraîner l'HMM avec des séquences provenant d'organismes distincts, on pourrait créer une annotation composite créée à partir de plusieurs annotations. Il serait possible, par exemple, de créer x annotations distinctes en entraînant le modèle avec x séquences d'entraînements provenant

d'organismes distincts. Si l'on désire une grande sensibilité, on pourrait générer l'annotation composite comme étant l'union des x annotations. Si au contraire on désire une grande spécificité, on pourrait produire l'annotation composite en prenant l'intersection des x annotations.

BIBLIOGRAPHIE

[1] S. F. Altschul et BW Erickson. Significance of nucleotide sequence alignments : a method for random sequence permutation that preserves dinucleotide and codon usage. *Mol Biol Evol*, 2(6):526–538, 1985.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers et D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, October 1990. ISSN 0022-2836.

[3] Leonard E. Baum, Ted Petrie, George Soules et Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970. ISSN 00034851.

[4] Ingrid Berg, Rita Neumann, Hakan Cederberg, Ulf Rannug et Alec J. Jeffreys. Two modes of germline instability at human minisatellite ms1 (locus d1s7) : Complex rearrangements and paradoxical hyperdeletion. *Cell*, 72:1436–1447, 2003.

[5] J. Bessereau. Transposons in c. elegans. *WormBook*, 2006.

[6] Harald Biessmann, James Mason, Kristian Ferry, Marie d'Hulst, Katrin Valgeirsdottir, Karen Traverse et Mary-Lou Pardue. Addition of telomere-associated het dna sequences "heals" broken chromosome ends in drosophila. *Cell*, 61(4):663 – 673, 1990. ISSN 0092-8674.

[7] Sebastian Bonhoeffer, Andreas V. M. Herz, Maarten C. Boerlijst, Sean Nee, Martin A. Nowak et Robert M. May. Explaining "linguistic features" of noncoding dna. *Science*, 271(5245):14–15, 1996. ISSN 00368075.

[8] Sebastian Bonhoeffer, Andreas V. M. Herz, Maarten C. Boerlijst, Sean Nee, Martin A. Nowak et Robert M. May. No signs of hidden language in noncoding dna. *Phys. Rev. Lett.*, 76(11):1977, Mar 1996.

[9] J. Brosius. Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica*, 1999.

[10] Brook Brouha, Joshua Schustak, Richard M. Badge, Sheila Lutz-Prigge, Alexander H. Farley, John V. Moran et Haig H. Kazazian. Hot L1s account for the bulk of retrotrans-

position in the human population. *Proceedings of the National Academy of Sciences of the United States of America*, 100(9):5280–5285, 2003.

[11] Michael Brown, Richard Hughey, Anders Krogh, I. Saira Mian, Sinsheimer Laboratories, Kimmen Sjolander et David Haussler. Using dirichlet mixture priors to derive hidden markov models for protein families. Dans *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology*, pages 47–55. AAAI Press, 1993.

[12] Jeremy Buhler. Provably sensitive indexing strategies for biosequence similarity search. Dans *RECOMB '02 : Proceedings of the sixth annual international conference on Computational biology*, pages 90–99, New York, NY, USA, 2002. ACM. ISBN 1-58113-498-3.

[13] Barbara Burwinkel et Manfred Kilimann. Unequal homologous recombination between line-1 elements as a mutational mechanism in human genetic disease. *Journal of Molecular Biology*, 277(3):513 – 517, 1998. ISSN 0022-2836.

[14] CA Chatzidimitriou-Dreismann, RM Streffer et D Larhammar. Lack of biological significance in the 'linguistic features' of noncoding DNA–a quantitative analysis. *Nucl. Acids Res.*, 24(9):1676–1681, 1996.

[15] Jean-Michel Claverie. Some useful statistical properties of position-weight matrices. *Computers & Chemistry*, 18(3):287 – 294, 1994. ISSN 0097-8485.

[16] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[17] N. L. Craig, R. Craigie, M. Gellert et A. M. Lambowitz. *Mobile DNA II.* Am. Soc. Microbiol, 2002.

[18] Nancy Craig, Orna Cohen-Fix, Rachel Green, Carol Greider, Gisela Storz et Cynthia Wolberger. Transposition via Target-Primed reverse transcription, juin 2002.

[19] Miklós Csurös, Laurent Noé et Gregory Kucherov. Reconsidering the significance of genomic word frequencies. *Trends in Genetics*, 23(11):543 – 546, 2007. ISSN 0168-9525.

[20] W. L. Delano. The pymol molecular graphics system, 2002.

[21] M. Dewannieux, C. Esnault et T. Heidmann. Line-mediated retrotransposition of marked alu sequences. *Nat Genet*, 35(1):41–8, 2003.

[22] C. Esnault, J. Maestre et T. Heidmann. Human line retrotransposons generate processed pseudogenes. *Nat Genet*, 24(4):363–7, 2000.

[23] Mark X. Geske, Anant P. Godbole, Andrew A. Schaffner, Allison M. Skolnick et Garrick L. Wallstrom. Compound Poisson approximations for word patterns under Markovian hypotheses. *J. Appl. Probab.*, 32(4):877–892, 1995. ISSN 0021-9002.

[24] S. Glinka, Lorenzo D. De et W. Stephan. Evidence of gene conversion associated with a selective sweep in drosophila melanogaster. *Mol Biol Evol*, 23(10):1869–78, 2006.

[25] A. Hagemann et N. Craig. Tn7 transposition creates a hotspot for homologous recombination at the transposon donor site. *Genetics*, 133(1):9–16, 1993.

[26] J. Han, S. Szak et J. Boeke. Transcriptional disruption by the l1 retrotransposon and implications for mammalian transcriptomes. *Nature*, 429(6989):268–74, 2004.

[27] K. Han, J. Lee, T. Meyer, J. Wang, S. Sen, D. Srikanta, P. Liang et M. Batzer. Alu recombination-mediated structural deletions in the chimpanzee genome. *Gene*, 3(10): 1939–49, 2007.

[28] K. Han, S. Sen, J. Wang, P. Callinan, J. Lee, R. Cordaux, P. Liang et M. Batzer. Genomic rearrangements by line-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Res*, 33(13):4040–52, 2005.

[29] E. Havecker, X. Gao et D. Voytas. The diversity of ltr retrotransposons. *Genome*, 2004.

[30] Jorja G. Henikoff et Steven Henikoff. Using substitution probabilities to improve position-specific scoring matrices. *Comput. Appl. Biosci.*, 12(2):135–143, 1996.

[31] T. Higashiyama, Y. Noutoshi, M. Fujie et T. Yamada. Zepp, a line-like retrotransposon accumulated in the chlorella telomeric region. *EMBO J*, 16(12):3715–23, 1997.

[32] N. Jameson, N. Georgelis, E. Fouladbash, S. Martens, L. Hannah et S. Lal. Helitron mediated amplification of cytochrome p450 monooxygenase gene in maize. *Plant Mol Biol*, 67(3):295–304, 2008.

[33] Minghui Jiang, James Anderson, Joel Gillespie et Martin Mayne. ushuffle : a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*, 9:192+, April 2008. ISSN 1471-2105.

[34] N. Jiang, Z. Bao, X. Zhang, S. Eddy et S. Wessler. Pack-mule transposable elements mediate gene evolution in plants. *Nature*, 431(7008):569–73, 2004.

[35] I. Jordan, I. Rogozin, G. Glazko et E. Koonin. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Gene*, 19(2):68–72, 2003.

[36] N. Juretic, D. Hoen, M. Huynh, P. Harrison et T. Bureau. The evolutionary fate of mule-mediated duplications of host gene fragments in rice. *Genome*, 15(9):1292–7, 2005.

[37] M. Kidwell. Transposable elements and the evolution of genome size in eukaryotes. *Genetica*, 115(1):49–63, 2002.

[38] J. Kleffe et M. Borodovsky. First and second moment of counts of words in random texts generated by Markov chains. *Comput. Appl. Biosci.*, 8(5):433–441, 1992.

[39] Bertrand Knebelmann, Lionel Forestier, Laurent Drouot, Susan Quinones, Christian Chuet, France Benessy, Juan Saus et Corinne Antignac. Splice-mediated insertion of an Alu sequence in the COL4A3 mRNA causing autosomal recessive Alport syndrome. *Hum. Mol. Genet.*, 4(4):675–679, 1995.

[40] Kazunori Kondo, Hachiro Inokuchi et Haruo Ozeki. The transposable element Tn3 promotes general recombination at the neighboring regions. *The Japanese journal of genetics*, 64(6):417–434, 1989.

[41] Andrzej K. Konopka et Colin Martindale. Noncoding dna, zipf's law, and language. *Science*, 268(5212):789, 1995. ISSN 00368075.

[42] Deininger P L et Batzer M A. Alu repeats and human disease. *Molecular genetics and metabolism*, 67(3):183–193, 1999.

[43] M.-Y. Leung, G. M. Marsh et Terence P. Speed. Over- and underrepresentation of short dna words in herpesvirus genomes. *Journal of Computational Biology*, 3(3):345–360, 1996.

[44] G. Lev-Maor, R. Sorek, N. Shomron et G. Ast. The birth of an alternatively spliced exon : 3' splice-site selection in alu exons. *Science*, 300(5623):1288–91, 2003.

[45] G. Levinson et G.A. Gutman. High frequencies of short frameshifts in poly-ca/tg tandem repeats borne by bacteriophage m13 in escherichia coli k-12. *Nucleic Acids Res.*, 15(13): 5323–5338, July 1987.

[46] Robert W. Levis, Robin Ganesan, Kathleen Houtchens, Leigh Anna Tolar et Fang miin Sheen. Transposons in place of telomeric repeats at a drosophila telomere. *Cell*, 75(6): 1083 – 1093, 1993. ISSN 0092-8674.

[47] J. Lingner, T. Hughes, A. Shevchenko, M. Mann, V. Lundblad et T. Cech. Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science*, 276(5312):561–7, 1997.

[48] N. M. Luscombe, J. Qian, Z. Zhang, T. Johnson et M. Gerstein. The dominance of the population by a selected few : power-law behaviour applies to a wide variety of genomic properties. *Genome Biol*, 3(8), July 2002. ISSN 1465-6914.

[49] M. Lynch et J. Conery. The origins of genome complexity. *Science*, 302(5649):1401–4, 2003.

[50] Bin Ma, John Tromp et Ming Li. PatternHunter : faster and more sensitive homology search . *Bioinformatics*, 18(3):440–445, 2002.

[51] H. S. Malik, S. Henikoff et T. H. Eickbush. Poised for contagion : Evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Research*, 10:1307–1318, 2000.

[52] R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C. K. Peng, M. Simons et H. E. Stanley. Linguistic features of noncoding dna sequences. *Phys. Rev. Lett.*, 73(23): 3169–3172, Dec 1994.

[53] Colin Martindale et Andrzej K. Konopka. Oligonucleotide frequencies in dna follow a yule distribution. *Computers & Chemistry*, 20(1):35 – 38, 1996. ISSN 0097-8485.

[54] G A Mitchell, D Labuda, G Fontaine, J M Saudubray, J P Bonnefont, S Lyonnet, L C Brody, G Steel, C Obie et D Valle. Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase : a role for Alu elements in human mutation.

Proceedings of the National Academy of Sciences of the United States of America, 88(3): 815–819, 1991.

[55] John V. Moran, Ralph J. DeBerardinis et Jr. Kazazian, Haig H. Exon Shuffling by L1 Retrotransposition. *Science*, 283(5407):1530–1534, 1999.

[56] T. Morrish, J. Garcia-Perez, T. Stamato, G. Taccioli, J. Sekiguchi et J. Moran. Endonuclease-independent line-1 retrotransposition at mammalian telomeres. *Nature*, 446(7132):208–12, 2007.

[57] K. Ohshima, M. Hattori, T. Yada, T. Gojobori, Y. Sakaki et N. Okada. Whole-genome screening indicates a possible burst of formation of processed pseudogenes and alu repeats by particular l1 subfamilies in ancestral primates. *Genome*, 2003.

[58] S Okazaki, H Ishikawa et H Fujiwara. Structural analysis of TRAS1, a novel family of telomeric repeat- associated retrotransposons in the silkworm, Bombyx mori. *Mol. Cell. Biol.*, 15(8):4545–4552, 1995.

[59] Frederic Paques, Wai-Ying Leung et James E. Haber. Expansions and Contractions in a Tandem Repeat Induced by Double-Strand Break Repair. *Mol. Cell. Biol.*, 18(4):2045–2054, 1998.

[60] M. Pardue et P. DeBaryshe. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annu Rev Genet*, 2003.

[61] M. Pardue, S. Rashkova, E. Casacuberta, P. DeBaryshe, J. George et K. Traverse. Two retrotransposons maintain telomeres in drosophila. *Chromosome Res*, 13(5):443–53, 2005.

[62] M. Pelé, L. Tiret, J. L. Kessler, S. Blot et J. J. Panthier. Sine exonic insertion in the ptpla gene leads to multiple splicing defects and segregates with the autosomal recessive centronuclear myopathy in dogs. *Human molecular genetics*, 14(11):1417–1427, June 2005. ISSN 0964-6906.

[63] O. Pickeral, W. Makaowski, M. Boguski et J. Boeke. Frequent human genomic dna transduction driven by line-1 retrotransposition. *Genome*, 10(4):411–5, 2000.

[64] C. Preston et W. Engels. P-element-induced male recombination and gene conversion in drosophila. *Genetics*, 144(4):1611–22, 1996.

[65] Bernard Prum, Francois Rodolphe et Elisabeth de Turckheim. Finding words with unexpected frequencies in deoxyribonucleic acid sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):205–220, 1995. ISSN 00359246.

[66] W. Reed. The double pareto-lognormal distribution - a new parametric model for size distribution, 2001.

[67] S. Robin et J. J. Daudin. Exact distribution of word occurrences in a random sequence of letters. *Journal of Applied Probability*, 36(1):179–193, 1999. ISSN 00219002.

[68] P. SanMiguel, B. Gaut, A. Tikhonov, Y. Nakajima et J. Bennetzen. The paleontology of intergene retrotransposons of maize. *Nat Genet*, 20(1):43–5, 1998.

[69] Sophie Schbath. Compound Poisson approximation of word counts in DNA sequences. *ESAIM Probab. Statist.*, 1:1–16 (electronic), 1995/97. ISSN 1292-8100.

[70] Y. Segal, B. Peissel, A. Renieri, Marchi M. de, A. Ballabio, Y. Pei et J. Zhou. Line-1 elements at the sites of molecular rearrangements in alport syndrome-diffuse leiomyomatosis. *Am J Hum Genet*, 64(1):62–9, 1999.

[71] Shurjo K. Sen, Kyudong Han, Jianxin Wang, Jungnam Lee, Hui Wang, Pauline A. Callinan, Matthew Dyer, Richard Cordaux, Ping Liang et Mark A. Batzer. Human genomic deletions mediated by recombination between alu elements. *The American Journal of Human Genetics*, 79(1):41 – 53, 2006. ISSN 0002-9297.

[72] G. Shalev et A. Levy. The maize transposable element ac induces recombination between the donor site and an homologous ectopic sequence. *Genetics*, 146(3):1143–51, 1997.

[73] A. F. A. Smit, R. Hubley et P. Green. Repeatmasker open-3.0, 1996-2004.

[74] W Stephan. Recombination and the evolution of satellite dna. *Genet Res*, 47(3):167–74, June 1986.

[75] Yanni Sun et Jeremy Buhler. Designing multiple simultaneous seeds for dna similarity search. Dans *RECOMB '04 : Proceedings of the eighth annual international conference on Resaerch in computational molecular biology*, pages 76–84, New York, NY, USA, 2004. ACM. ISBN 1-58113-755-9.

[76] Yanni Sun et Jeremy Buhler. Choosing the best heuristic for seeded alignment of dna sequences. *BMC Bioinformatics*, 7(1):133, 2006. ISSN 1471-2105.

[77] Jack W. Szostak, Terry L. Orr-Weaver, Rodney J. Rothstein et Franklin W. Stahl. The double-strand-break repair model for recombination. *Cell*, 33(1):25–35, 1983. ISSN 0092-8674.

[78] H. Takahashi, S. Okazaki et H. Fujiwara. A new family of site-specific retrotransposons, sart1, is inserted into telomeric repeats of the silkworm, bombyx mori. *Nucleic Acids Res*, 25(8):1578–84, 1997.

[79] S. Tsubota, D. Rosenberg, H. Szostak, D. Rubin et P. Schedl. The cloning of the bar region and the b breakpoint in drosophila melanogaster : evidence for a transposon-induced rearrangement. *Genetics*, 122(4):881–90, 1989.

[80] Elisabetta Ullu et Christian Tschudi. Alu sequences are processed 7SL RNA genes. *Nature*, 312:171–172, 1984.

[81] R. Varon, R. Gooding, C. Steglich, L. Marns, H. Tang, D. Angelicheva, K. Yong, P. Ambrugger, A. Reinhold, B. Morar, F. Baas, M. Kwa, I. Tournev, V. Guerguelcheva, I. Kremensky, H. Lochmüller, A. Müllner-Eidenböck, L. Merlini, L. Neumann, J. Bürger, M. Walter, K. Swoboda, P. Thomas, Moers A. von, N. Risch et L. Kalaydjieva. Partial deficiency of the c-terminal-domain phosphatase of rna polymerase ii is associated with congenital cataracts facial dysmorphism neuropathy syndrome. *Nat Genet*, 35(2): 185–9, 2003.

[82] R. Vervoort, R. Gitzelmann, W. Lissens et I. Liebaers. A mutation (ivs8+0.6kbdeltc) creating a new donor splice site activates a cryptic exon in an alu-element in intron 8 of the human beta-glucuronidase gene. *Gene*, 103(6):686–93, 1998.

[83] W. Wei, N. Gilbert, S. Ooi, J. Lawler, E. Ostertag, H. Kazazian, J. Boeke et J. Moran. Human l1 retrotransposition : cis preference versus trans complementation. *Cell*, 21(4): 1429–39, 2001.

[84] Yong-Li Xiao, Xianggan Li et Thomas Peterson. Ac Insertion Site Affects the Frequency of Transposon-Induced Homologous Recombination at the Maize p1 Locus. *Genetics*, 156(4):2007–2017, 2000.

[85] Z. Yu, S. Wright et T. Bureau. Mutator-like elements in arabidopsis thaliana. structure, diversity and evolution. *Genetics*, 156(4):2019–31, 2000.

[86] L. Zhang, H. Lu, W. Chung, J. Yang et W. Li. Patterns of segmental duplication in the human genome. *Mol Biol Evol*, 22(1):135–41, 2005.

[87] George K. Zipf. *Human Behavior and the Principle of Least Effort.* Addison-Wesley (Reading MA), 1949.